DOCUMENT RESUME

ED 039 584 CG 005 379

AUTHOR	Miller, George E.; And Others
TITLE	The Orthopaedic Training Study. Final Report.
INSTITUTION	Illinois Univ., Chicago. Center for the Study of Medical Education.
SPONS AGENCY	Public Health Service (DHEW), Arlington, Va. Bureau of Health Manpower.
PUB DATE	[68]
NOTE	339p.
EDRS PRICE	EDRS Price MF-$1.50 HC-$17.05
DESCRIPTORS	Evaluation Criteria, Evaluation Methods, *Evaluation Techniques, *Physicians, *Professional Education, *Professional Training, *Training

ABSTRACT

A four year study was initiated to systematically improve the certification procedures of the American Board of Orthopaedic Surgery. Consequently, the immediate research aim was the development of more valid and reliable techniques in assessing professional competence in orthopedics. A definition of professional competence was reached through utilization of the Critical Incidence Technique, which resulted in a list of objectives to be used as a guide in developing instructional programs and evaluative instruments. Old techniques for assessing competence were analyzed and found inadequate. Specific instruments developed in the project are described in detail. They include: (1) rating forms; (2) tests and test manuals; (3) forms for profile scoring and reporting of test results; and (4) observational forms for process analysis of tests. Analytical, simulation, and observational approaches were employed in the design of the instruments. Intermediate and long-term outcomes of the study are discussed and a proposed followup study described. (TL)

DOCUMENT FILMED FROM BEST AVAILABLE COPY

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED0 39584

THE ORTHOPAEDIC TRAINING STUDY

FINAL REPORT

Harold G. Levine, M.P.A.
Project Officer

Christine McGuire, M.A.
Project Director

George E. Miller, M.D. and Carroll Larson, M.D.
Co-Principal Investigators

The research upon which this report is based was performed pursuant to Research Grant No. PM 00014 Bureau of Health Manpower, Public Health Service, Department of Health, Education and Welfare

CG005379

Center for the Study of Medical Education
University of Illinois College of Medicine

RESEARCH STAFF

Professional Staff

George E. Miller, M.D., Principal Investigator, Director, Office of Research in Medical Education, University of Illinois College of Medicine.

Carroll B. Larson, M.D., Co-Principal Investigator, University Hospitals, Iowa City, Iowa.

Christine McGuire, M.A., Project Director, Chief-Evaluation Studies Section, Office of Research in Medical Education, University of Illinois College of Medicine.

Harold G. Levine, M.P.A., Project Officer, Senior Associate, Office of Research in Medical Education, University of Illinois College of Medicine.

LeRoy W. Nattress, Director, Office of Education and Evaluation, American Board of Orthopaedic Surgery.

John R. Noak, Ed.D., Associate, Office of Research in Medical Education, University of Illinois College of Medicine.

Brian E. Huncke, M.D., Associate and Clinical Professor, University of Illinois.

Jo Miller, M.D., Chairman, Department of Orthopaedics, Montreal General Hospital.

Consulting Staff

Samuel A. Barley, M.D., Akron City Hospital, Akron, Ohio.

Warren G. Stamp, M.D., University of Virginia, Charlottesville, Virginia.

L. H. Paradies, M.D., Southwestern Medical School, Dallas, Texas.

Robert Johnston, M.D., United States Air Force Academy, Boulder, Colorado

Reginald Cooper, M.D., State University of Iowa Medical School,
Iowa City, Iowa.

Floyd E. Bliven, M.D., Medical College of Georgia, Augusta,
Georgia.

Russell Grunsten, M.D., Tulane University Medical School,
New Orleans, Louisiana.

Douglas W. McKay, M.D., Shriners Hospital for Crippled Children,
Shreveport, Louisiana.

Secretarial and Technical Staff

E. O. Branson, Administrative Assistant
Richard Carter, Data Processing Analyst
Vonita Curbow, Secretary
Gale Kappe, Secretary

ACKNOWLEDGMENT

The members of the American Board of Orthopaedic Surgery served in an advisory capacity throughout this study. All their efforts in behalf of the study and their assistance to the Research Staff are gratefully acknowledged.

TABLE OF CONTENTS

List of Staff i
Acknowledgements iii

PREFACE

Chapter I Introduction and Overview 1

SECTION ONE: PRELIMINARY STEPS

Chapter II Definition of Professional Competence: The Critical Incident Study 9

Chapter III Analysis of the Techniques Currently Employed 15

SECTION TWO: DEVELOPMENT AND ANALYSIS OF NEW EVALUATION TECHNIQUES

Chapter IV Statement of the Problem 36

Chapter V Observation Forms 43

Chapter VI Written Simulation Exercises 59

Chapter VII Oral Examination 80

Chapter VIII The Multiple Choice Questions 116

Chapter IX A Model for the Evaluation of Competence: A Synthesis of the New Techniques 128

Chapter X The Application of Profile Technique to the Problems of Certification 136

SECTION THREE: CURRENT STATUS AND PROJECTED NEXT STEPS

Chapter XI Outcomes of the Orthopaedic Training Study 148

Chapter XII A Look to the Future 157

APPENDIX

Appendix 1 Determination of Objectives 161

Appendix 2 A Taxonomy of Intellectual Processes 169

Appendix 3 Working Papers: Task Force on Written Written Examinations 173

Appendix 4 Observation Form and Instruction for Use 177

Appendix 5 Resident Evaluation Form (Experimental Form for Research Purposes) 181

Appendix 6 Resident Evaluation Form (Standard Form for Regular Administration with Orthopaedic In-Training Examination) 185

Appendix 7 Candidate Evaluation Form 187

Appendix 8 Distribution of Ratings, Candidate Evaluation Form, for 1968 Final Certification Examination 195

Appendix 9 Rating Form: Evaluation of Surgical Skill 197

Appendix 10 Oral Examination: Form for Rating Patient Interviews 201

Appendix 11 Oral Examination: 1967 Form for Rating Simulated Patient Management Conference 205

Appendix 12 Oral Examination: 1966 Form for Rating Simulated Patient Management Conference 209

Appendix 13 Oral Examination: 1968 Standard Rating Form 213

Appendix 14 Multiple Regression Analysis, 1966 In-Training Examination 215

Appendix 15 Multiple Regression Analysis, 1968 Certifying Examination 221

Appendix 16 Instructions to Examiners for Oral Examinations 227

Appendix 17 Instructions to Candidates 243

Appendix 18 Oral Examiner Questionnaire 249

Appendix 19 Candidate Questionnaire 255

Appendix 20 Simulated Patient Management Conference Case Description No. 35 261

Appendix 21 System of Content Classification of Examination Materials 263

Appendix 22 Examination Blueprint 265

Appendix 23 Working Papers: Special Meeting of Examination Committee 269

Appendix 24 Working Papers: Task Force on Oral Examinations 281

Appendix 25 Working Papers: Task Force on Weighting and Scoring 291

Appendix 26 Setting Standards of Competence: The Minimum Pass Level 305

Appendix 27 Working Papers: Meeting of Task Force Chairmen and Examination Committee 317

Appendix 28 Process Analysis of Oral Patient Management Problems, Interpretive Skills and Role-Playing Examinations 325

Appendix 29 Chronological List of Joint Activities of American Board of Orthopaedic Surgery and Center for the Study of Medical Education 337

Appendix 30 A Proposal for a Ten-year Follow-up 343

PREFACE

CHAPTER I:

INTRODUCTION AND OVERVIEW

In June, 1964 the Center for the Study of Medical Education at the University of Illinois College of Medicine embarked on a four-year joint study with the American Board of Orthopaedic Surgery in an attempt to develop improved methods of assessing competence in the field of orthopaedic surgery. The study was supported by the Bureau of State Services of the Public Health Service as one approach to obtaining better utilization of health manpower by means of increased flexibility and efficiency in the training of health professionals.

The following were deemed essential to the accomplishment of this goal:

1. Precise identification of the components of professional competence in the field
2. Development of valid and reliable technique of assessing these components of competence
3. Identification of variations in the patterns of competence and differential rates in their achievement associated with variations in training programs

Consequently, the first stage of the investigation was devoted to a critical incident study[1] of the essential performance requirements for orthopaedic surgeons. The resulting definition of competence included 9 major categories (such as "Skill in Gathering Clinical Information," "Competence in Developing a Diagnosis," and "Effectiveness of Physician-Patient Relationship") and 94 subcategories of behavior.[2]* This definition of competence served to direct all subsequent stages of the study.

At the same time that the critical incident study was being conducted a task force of orthopaedic surgeons under the direction of the Center staff analyzed the behaviors sampled by the written examinations currently in use by the Board and concluded that most questions required primarily the ability to recall isolated fragments of information. Analogous observational study of the oral examinations yielded similar results.

The next stage of the study was therefore devoted to the improvement of conventional techniques and to the development of new ones designed to assess the other components of competence identified in the critical incident study. During this period therefore, in addition to improving the quality of multiple choice exercises, the following new techniques were developed:

---

* See Chapter II and Appendix 1.

1. Written simulation exercises
2. Oral simulation exercises utilizing role-playing techniques
3. Oral exercises testing complex cognitive abilities
4. Rating forms for evaluating habitual performance
5. Rating forms for evaluating specific abilities in single observations

To study the validity and reliability of these techniques several forms of these examinations were developed and administered to the following populations of examinees:

1. A final orthopaedic certification examination (OCE) was administered to 4 populations composed of candidates for certification who had completed their residency and were currently in orthopaedic practice;
2. A prerequisite certifying examination (OCE-I) was administered to 2 populations composed of candidates for certification who were in their last year of residency training;
3. An In-Training Examination (ITE) was administered by the American Academy of Orthopaedic Surgeons, for diagnostic purposes to 3 populations composed of virtually all residents currently in training;
4. The written simulation exercises from one of the final orthopaedic certification examinations (OCE) was administered for experimental purposes to all Board examiners.

The following forms of the above listed examinations were analyzed to obtain reliability and validity data on both the new and the more conventional techniques:

The January 1965, 1966, 1967 and 1968 final Orthopaedic Certification Examinations; the May 1965 and 1966 prerequisite Certification Examinations--1; and the November 1965, 1966 and 1967 In-Training Examinations.

The reliability of the written forms was assessed by analysis of internal consistency; the reliability of the orals was assessed by analysis of inter-rater agreement and by correlation of alternate forms; the reliability of ratings of habitual performance was assessed by analysis of interrater agreement.

The content validity of the examinations was assessed by process analysis.[3] Their construct validity was assessed by testing hypotheses regarding the relationship between performance on the examination and such factors as age, experience, and practice settings; additional data on construct validity was obtained from various correlational and factor analytic studies. Concurrent validity of the tests was assessed by correlational and multiple regression analysis of the relationships between test scores and supervisor's ratings. Predictive validity of the examinations is to be assessed in a 10-year follow-up study.*

The detailed results of these analyses are discussed in Chapters IV through X below; however, the initial findings can be briefly summarized as follows:

1. Numerous skills and abilities are requisite to competence in orthopaedic surgery and the correlation between certain of them is relatively low, e.g., "Surgical Skill" versus "Ability to Relate to Patients."

2. Each of the new evaluation techniques (as listed earlier) appears to measure certain independent aspects of competence not assessed by other techniques.

3. These newer and more complex techniques tend to be less reliable than conventional "objective" (ie., multiple choice) techniques primarily because fewer independent samples of behavior can be obtained in a given time; however, they are considerably more reliable than techniques which depend upon ratings (either of habitual behavior or of single incidents), and their reliability can be increased by pooling response data from a number of techniques and from repetitions of one technique to arrive at a composite score.

4. The concurrent validity of this composite score appears to be substantially higher than that of scores obtained from conventional techniques of testing and of pooling response data.

---

* See Chapter XII and Appendix 30

On the basis of these results the study staff recommended restructuring the entire certification procedure. These recommendations and the supporting data were reviewed by a group of special Task Forces appointed by the Board. The consequent modifications in the certification procedures, as adopted by the Board, were first fully implemented in the January, 1968 Final Certification Examination. Table 1 summarizes the differences between that examination and the one administered in 1964, the year just prior to the initiation of the Orthopaedic Training Study.

With the elaboration of a behavioral definition of competence in orthopaedic surgery, the development of more valid and reliable techniques for assessing these competencies and the incorporation of these techniques in an integrated certification system, it has now become possible to identify more precisely the relationship between variations in training programs and differential rates and patterns of achievement. The present study, therefore, serves both as one type of model for professional self study and as the indispensable prerequisite for a further study of methods of increasing the efficiency and effectiveness of training in this specialty. The first section of the present study report is devoted to a discussion of the rationale and findings of the prior analyses required in the development of new methods of evaluating professional competence in this specialty; the second section contains a description of each new assessment technique developed during this study, the methods employed in analyzing the validity and reliability of each, and the findings from each such analysis; the third section summarizes the methods used in implementing new certification procedures based on the research findings, and the outlined plans for subsequent study.

---

[1] John C. Flanagan, "The Critical Incident Technique," Psychological Bulletin. July, 1954. Vol. 51, No. 4, pp. 327-358.

[2] J. Michael Blum and Robert Fitzpatrick, Critical Performance Requirements for Orthopaedic Surgery, (2 volumes) Pittsburgh, Pa. American Institutes for Research, 1965.

[3] Christine McGuire, "A Process Approach to the Construction and Analysis of Medical Examinations," The Journal of Medical Education, Vol. 38, No. 7, July, 1963.

TABLE 1: TWO SYSTEMS OF BOARD CERTIFICATION

Aspect	January 1964	January 1968
Requirements for eligibility	Completion of an approved residency	Same as 1964
	Completion of two years of practice or its equivalent	Completion of one year of practice on its equivalent
	Satisfactory completion of Part I examination taken at end of residency	Requirement eliminated
	Letters of recommendation from chief of service and current colleagues	Same as 1964 plus submission of standardized Candidate Rating Form by training chiefs
Method of preparation f examination	Various subject matter parts of the written assigned to different members of the Examination Committee for development; All materials reviewed by entire Committee in 2-3 day meeting.	Detailed set of specifications with respect to content and process, established by the Examination Committee and approved by Board; Preparation of materials (both written and oral) to meet these specifications assigned to various task forces; Preliminary form of written test administered to entire Board; data from this initial try out reviewed by Examination Committee as basis for composing test in final form; All materials reviewed by Examination Committee

TABLE 1: (Cont'd)

Aspect	January 1964	January 1968
Standards for Certification	Oral: 75 or better on every oral Written: Not lower than one standard deviation below the mean of a subgroup composed of all graduates of U.S. medical schools who had no previous failures on the Board Examination.	Achievement of the pre-established "minimum passing level "on the weighted total score and on the Recall and Problem-Solving factors; and achievement of "marginal level" on at least 2 of the 4 factors including either the Recall OR Problem-Solving factor.
Feedback to Candidates	Scores reported as Pass-Fail on each technique: written or oral and, within the orals, on each discipline.	Scores reported as overall Pass-Fail together with a report of deficiency on any factor.
Training of Examiners	Informal induction in an apprentice-like system with general guidelines explained by subject advisers in an evening pre-session and general postmortem by panel advisers in an evening post session.	Formal 1-2 day workshops on test construction for authors of written test. Formal 1-2 day training sessions on administering and scoring oral examinations.
Scoring of Examination	Separate scores (based on scale of 100) derived for the written and each oral.	One or more of the following sub-scores derived from each test: Recall, Observation and Interpretation, Problem-solving, Ability to Communicate with patients and colleagues; these sub-scores converted to a common 12-point scale and combined across tests to yield an overall score on each factor named above and a weighted total score on all factors combined.

TABLE 1 : (Cont'd)

Aspect	January 1964	January 1968
Method of preparation of examination	Orals in specific subjects developed by individual examiners in accord with general guidelines, no prior review of these materials	Standardized cases and related materials for orals prepared by task forces and reviewed by the Examination Committee
Method of Examination	2 hour multiple choice plus 2 1/2 hour oral consisting of 1/2 hour oral quiz in each of 5 subject fields: adult, children's, trauma, anatomy and pathology; variously designed questions and case materials individually prepared and selected by examiners.	2 hour multiple choice test plus 1 hour written simulated patient management problems 3 half-hour orals on patient management problems in adult, children's and trauma, utilizing standardized cases administered by trained examiners, plus 1 half-hour oral on interpretation of X-rays and histologic materials, using standardized cases administered by trained examiners, plus 1 half-hour oral simulating various physician-patient and physician-colleague encounters, utilizing standardized case materials administered by trained examiners.
	Most questions in the oral and written designed to assess recall of information	Most questions in both orals and written designed to assess skills other than recall

SECTION ONE

PRELIMINARY STEPS

CHAPTER II

DEFINITION OF PROFESSIONAL COMPETENCE: THE CRITICAL INCIDENT STUDY

Obtaining a meaningful statement of behavioral objectives in a form suitable for the direction of medical education has often proved difficult because those developed by subject matter specialists are typically so vague (e.g., to produce physicians who are good at critical thinking) as to provide little guidance to those responsible for program planning. Secondly, the objective sought by one instructor may not be shared by his colleagues and there is little basis for choice between differing views. What is needed, therefore, is a list of objectives which are specific enough to use as a guide in developing instructional programs and evaluation instruments, and which are at the same time general enough to be acceptable to all those responsible for the educational program. One method of defining objectives which meets these criteria is that developed during World War II by Flanagan and his associates in an attempt to improve the efficiency of pilot training.[1] Very briefly this approach, known as the "Critical Incident Technique," consists in collecting descriptions of several thousand specific incidents involving effective or ineffective performance by individuals in training. These incidents are reviewed and classified in empirically derived categories that describe the essential element of behavior that seems to account for the effective or ineffective performance. In the present study over 1700 such incidents involving effective and ineffective performance of orthopaedic surgeons were collected from the almost 3,000 members of the specialty contacted during the first year of the study. The number and sources of the incidents collected are shown in Table 2. These were classified into the types of categories listed in Exhibit I. (For a complete list of all sub-categories see Appendix 1)

The types of incidents collected and the nature of the categories derived from them are illustrated in Exhibit II.[2] From the examples given it is obvious that the specific incidents can be used to define the components of competence in behavioral terms. The critical incident technique therefore provides the optimum in specificity. Furthermore, since the categories are obtained empirically, the technique provides a basis for consensus concerning those aspects of professional competence that should be evaluated. Finally, it makes explicit the categories which qualified experts who are broadly representative of the specialty actually use to make value judgments about performance.

TABLE 2

A SUMMARY OF ORTHOPAEDISTS CONTACTED AND CRITICAL INCIDENTS COLLECTED

Population Sampled	_N_ Contacted	_N_ Incidents Returned	Incidents Per Contact
Phase 1			
Regional Attendees	240*	124	.517*
Regional Recommendees	136	93	.684
Examiners	155	108	.697
American Orthopaedic Association	150*	66	.440*
Canadian Orthopaedic Association	117	10	.085
Residents (through 280* Residency Directors)	1200*	224	.187*
Phase 2: Meetings			
Iowa City	25*	9	.360*
New Orleans	14	47	3.357
Cleveland	43	111	2.581
Portland	19	74	3.895
Washington	35*	86	2.457*
Phase 2: Mail			
Candidates	311	516	1.659
1963 Certificands	193	82	.425
1964 Certificands	243	211	.868
TOTALS	2881*	1761	.611*

*Approximation

Exhibit I:

Major Categories and Illustrative Sub-Categories Defining Critical Performance Requirements for Orthopaedic Surgery

I. Skill in Gathering Clinical Information
  A. Eliciting Historical Information
  B. Obtaining Information by Physical Examination
  C. Etc...
II. Effectiveness in Using Special Diagnostic Methods
  A. Obtaining and Interpreting X-Rays
  B. Obtaining Additional Information by Other Means
  C. Etc...
III. Competence in Developing a Diagnosis
  A. Approaching Diagnosis Objectively
  B. Recognizing Condition
  C. Etc...
IV. Judgment in Deciding on Appropriate Care
  A. Adapting Treatment to the Individual Case
  B. Determining Extent and Immediacy of Therapy Needs
  C. Etc...
V. Judgment and Skill in Implementing Treatment
  A. Planning the Operation
  B. Making Necessary Preparations for Operating
  C. Modifying Operative Plans According to Situation
  D. Etc...
VI. Effectiveness in Treating Emergency Patients
  A. Handling Patient
  B. Performing Emergency Treatment
  C. Etc...
VII. Competence in Providing Continuing Care
  A. Attention Post-Operatively
  B. Monitoring Patient's Progress
  C. Etc...
VIII. Effectiveness of Physician-Patient Relationship
  A. Showing Concern and Consideration
  B. Relieving Anxiety of Patient and Family
  C. Etc...
IX. Accepting Responsibility for Welfare of Patient
  A. Accepting Responsibility for Welfare of Patient
  B. Recognizing Professional Capabilities and Limitations
  C. Relating Effectively to Other Medical Persons
  D. Etc...

EXHIBIT II

IV. Judgment in Deciding on Appropriate Care

D. Modifying operative plans according to situation

2. Improvising with implements and materials. The orthopaedists can use materials in makeshift or innovative fashion.

EFFECTIVE

Situation:	During an open reduction of a compound fractured radius, part of the bone was so comminuted that it could not be reapproximated so as to restore length.
Experience:	This was a Board certified physician.
Action:	Took graft from pelvis and shaped it to resemble the missing bone. Placed Rush Rod down center of graft and replaced it in forearm.
Why Effective:	This idea was original, not preconceived and answered the problem at hand.
Less Effective:	Accept reduction.

VIII. Effectiveness of Physician-Patient Relationship

B. Relieving anxiety of Patient and Family

2. Explaining condition, treatment, proposals or complication. The orthopaedist informs the patient and/or family, in terms which they can understand, of the progress of therapy.

INEFFECTIVE

Situation:	Child with Legg-Perthe's disease.
Experience:	This was a Board certified physician with approximately three years of post-residency practice.
Action:	Failed to discuss thoroughly with the parents the type of treatment being given and the reasons for it. Parents misunderstood completely the function of the brace and the build-up on the opposite shoe, and thought that the surgeon was treating the wrong leg and sued him for malpractice.
Why Ineffective:	Surgeon should be careful to explain thoroughly his treatment and the reasons for it wherever possible.

The technique does, however, have certain weaknesses which should be noted. In general, it can be reasonably stated that the main function of professional education is to prepare people to perform a certain role in society. If one uses the critical incident technique to define this role, one risks two types of errors: First, members of a profession may have a narrower view of the role of a professional than do his clients or colleagues. Since patients, other physicians and paramedical personnel are also concerned with the roles played by orthopaedists, a critical incident study that included these groups might uncover areas of competence which the orthopaedists ignore. This is an important criticism and is particularly serious if the sample from whom incidents are collected is either very small or unusually homogenous. The second potential source of errors is attributable to possible bias on the part of the observer who records the incident. While this problem is mitigated by the fact that thousands of incidents are collected from hundreds of individuals the technique does not eliminate group biases characteristic of an entire profession.

In addition to these obvious sources of error it should be observed that there are a number of problems in developing behavioral objectives which are not solved by the critical incident technique. First, it provides no guidance regarding priorities since the number of incidents recorded in any category reflects the incidence of the behavior not the importance of the behavior. (See Table 3)

TABLE 3
NUMBER OF INCIDENTS REPORTED
IN EACH CATEGORY

Category		Number
1.	Skill in Gathering Clinical Information	59
2.	Effectiveness in Using Special Diagnostic Methods	60
3.	Competence in Developing a Diagnosis	109
4.	Judgment in Deciding on Appropriate Care	416
5.	Judgment and Skill in Implementing Treatment	297
6.	Effectiveness in Treating Emergency Patients	72
7.	Competence in Providing Continuing Care	84
8.	Effectiveness of Physician-Patient Relationship	125
9.	Accepting Responsibilities of a Physician	523

In this study this limitation of the technique was obviated by the decision that competence in orthopaedics is multi-dimensional and that candidates should therefore meet minimally standards in all behavioral categories; excellence in one area could not compensate for deficiency in another. Thus it was unnecessary to decide whether, for example, surgical skill is more important than diagnostic ability; a competent orthopaedist must meet minimal satisfactory standards in each area of competency. However, this decision, in itself, created other practical problems since it was necessary to reduce the 94 categories of behavior derived from the critical incident study into some manageable number for purposes of professional assessment.

The categories of performance that were finally chosen were logical but arbitrary groupings and distillations of the original 94. It might be that a different team of evaluators would have developed a different set of categories. As finally agreed upon by the Board of Orthopaedic Surgery and the research staff they consisted in the following groupings: (1) recall of basic information; (2) observation and interpretation of relevant data; (3) skill in problem-solving; and (4) ability to relate effectively to patients and colleagues, (5) surgical skill and (6) moral and ethical qualities. These six components of competence, as defined and specified by the critical incident study of performance in orthopaedic surgery, have served to direct all subsequent steps in the research project.

---

[1] John C. Flanagan, "The Critical Incident Technique," Psychological Bulletin. July, 1954. Vol. 51, No. 4, pp. 327-358.

[2] J. Michael Blum and Robert Fitzpatrick, Critical Performance Requirements for Orthopaedic Surgery, (2 volumes) Pittsburgh, Pa. American Institutes for Research, 1965.

CHAPTER III

ANALYSIS OF THE TECHNIQUES CURRENTLY EMPLOYED

Once behavioral objectives are developed, it is necessary to review existing evaluation instruments to determine how well they sample the critical behaviors. This task is not easy because most test exercises are abstractions which sample elements of what test constructors believe are important prerequisites to effective behavior. Thus if one wishes to evaluate a chemist, one ordinarily does not observe him in his laboratory since such assessments are usually both impractical and unreliable, but instead one develops a test which, in theory, samples the elements of effective performance as a chemist.

It is generally recognized that two assumptions are involved in assessments of this kind. The first assumption is that some exercises measure the recall of information which is a necessary, but not sufficient, requirement for effective performance. This assumption is rarely tested but there is a great deal of evidence that it may be dubious, first, because many practitioners of a profession do not depend wholly upon their memories but use handbooks and reference works and, second, because some information demanded by many tests is so esoteric that it is doubtful if the information is needed for any conceivable purpose.

The second assumption is that some of the exercises require behavior which closely imitates that required for effective performance. Thus an accountant may be given an exercise which closely approximates reality. The testing of this assumption, is, therefore, an important aspect in assaying the effectiveness of any examination.

It was the necessity to test these assumptions that led Bloom and his associates to develop the system for analyzing and classifying test exercises described in the Taxonomy of Educational Objectives Handbook I, The Cognitive Domain.[1] This taxonomy is based on the premise that some test exercises demand only the recall of isolated bits of information while others require the examinee to demonstrate his ability to apply information to the solution of problems. The levels of intellectual process therein outlined are generalized to apply, in principle, to all educational levels and are not specific to medical education. For this reason, a modification of the Bloom taxonomy, developed by the Committee on Student Appraisal of the University of Illinois College of Medicine and adapted to medical education (See Appendix 2) was employed to analyze both the written and oral examinations prepared by the American Board of Orthopaedic Surgery.

EXHIBIT III: A Taxonomy of Intellectual Processes

Level 1:	Recall and recognition of information
Level 2:	Selection of a relevant generalization to explain specific phenomena
Level 3:	Problem solving of a familiar type requiring simple interpretation of data or the application of a single principle or a standard combination of principles to a situation of a familiar type
Level 4:	Problem solving of an unfamiliar type requiring analysis of data or the application of a unique combination of principles to solve a problem of a novel type
Level 5:	Evaluation of a total situation
Level 6:	Synthesis of a variety of elements of knowledge into an original and meaning whole

Analysis of Written Examination

Process Analysis

Utilizing the hierarchically ordered classification system shown in Exhibit III a Task Force consisting of four orthopaedic surgeons and two test specialists met and independently rated each question in the January 1964 Orthopaedic Certification Examination and the May 1964 Orthopaedic Certification Examination Part I according to the highest intellectual process which the "typical" candidate would need to employ in responding to the 422 questions comprizing these two examinations.* As shown in Table 4 there was complete agreement among all four raters on about half the items and substantial disagreement on about one-fourth. Further analysis revealed that the disagreements were attributable to the following factors:

---

* See Appendix 3 for working papers and instructions to this Task Force.

1. Most of the disagreements occurred in the first set of items which were independently classified by these experts, at a point when they were as yet unfamiliar with the general approach and the specific system of classification.

2. The hierarchical nature of the system of classification created problems; some raters were inclined to rate questions according to the predominant process involved, rather than, as previously agreed, according to the highest level required.

3. Many items were part of a series all of which were based on data presented in the form of a clinical situation. Some task force members initially adopted the practice of classifying all items in such a group at the highest level required at any point in the analysis of the situation, whereas other raters followed the practice of according this very high rating to only one or two questions (and the specific items so classified varied from expert to expert) on the ground that once these questions had been answered other items in the group (e.g. about therapy, next diagnostic steps etc.) in involved only recall or generalization about the disease entity described.

4. On other items there was either uncertainty or difference of opinion about the nature of the experience most candidates would have had with a specific type of clinical problem described and hence about the process the "typical" candidate would need to employ.

5. Finally some items were classified differently by the several members of the Task Force because the formulation of the question or the alternatives presented difficulties of interpretation that were artifacts of the wording and not inherent in the question being posed.

Disagreements arising from the first three sources noted above and some arising from the fourth were readily resolved in the group sessions following the independent rating of the first 233 questions. Principles of classification evolved in these discussions clarified the categories and appeared to reduce potential disagreements in subsequent classifications.

TABLE 4: RATER AGREEMENT IN 1964
PROCESS ANALYSIS OF MULTIPLE CHOICE QUESTIONS

Test		Number of Items				
		With Identical Ratings by ALL Raters*	With Identical Ratings by Three of Four Raters	With Ratings Divided Between Two Adjacent Levels	Other	Total
OCE - January 1964		44	26	19	61	150
OCE I - Section I - May 1964		59	#	11	13	83
OCE I - Section II - May 1964		50	#	26	11	87
OCE I - Section III - May 1964		57	#	21	5	83
OCE I - Section IV - May 1964		7	#	2	10	19
TOTAL	Number	217	26	79	100	422
	% of Ratings	51.4%	6.2%	18.7%	23.7%	100%

* Among four raters on Part II and three on Part I

# Not applicable

TABLE 5: FINAL RESULTS OF 1964
PROCESS ANALYSIS OF MULTIPLE CHOICE QUESTIONS

Level	Number of Independent Ratings at Each Level								
	Prior to Discussion of Discrepancies							After Reconciliation of Discrepancies	
	OCE	OCE--I, May, 1964							
	January 1964	Section I	Section II	Section III	Section IV	Total No.	%	Total No.	%
6: Synthesis	0	0	0	0	0	0	0	0	0
5: Evaluation	51	0	1	0	0	52	3.7	11	0.4
4: Unfamiliar Problem-Solving	9	23	19	15	0	66	4.7	42	3.0
3: Familiar Problem-Solving	139	14	52	28	34	267	18.8	318	22.5
2: Generalization	47	38	2	3	3	93	6.6	82	5.8
Recall	350	174	179	202	18	923	65.2	924	65.2
Omit	4	0	8	1	2	15	1.1	39	2.8
Tot	600	249	26	249	57	1416	100.1	14	100.1

The final results, summarized in Table 5, indicate that:

1. There was substantial agreement among raters on 75% of the 422 items reviewed;

2. Over half the items were unanimously believed to require only recall of information;

3. Fewer than 25% of the items were thought by any expert to involve even simple interpretation of data, application of principles or evaluation;

4. Only four of the items were thought by any rater to involve evaluation of a total situation;

5. No item was thought to require synthesis.

Statistical Analysis

The results of the process analysis reported above were further substantiated by subsequent factor analytic studies of the 1966 and 1968 final Orthopaedic Certification Examinations. In addition to conventional written and oral components, these examinations included new assessment techniques deliberately designed to evaluate abilities not adequately assessed by the traditional methods. If the rationale used in analyzing the old techniques and developing the new is correct then it is logical to predict that the old and new techniques would load on different factors. As reported in Table 6 and 7 such is indeed the case: Both the 1966 and 1968 examinations show a similar factor pattern in which the conventional techniques have high loadings on one factor and most of the new techniques have high loadings on other factors.

Data on the concurrent validity of the several techniques also supports the conclusions based on the process analysis of the written examinations. Repeatedly such studies indicate that the score on the multiple choice component of any orthopaedic examination is the best predictor of supervisors' ratings of residents on such factors as "recall of factual information" or "ability to gather information," but that it is less valid than other techniques as a predictor of their ratings on such factors as "problem-solving skill." (The detailed discussion of the construct and concurrent validity of the multiple choice technique is included in Chapter VIII)

In summary, there is little doubt that the multiple choice technique as it was employed by the American Board of Orthopaedic Surgery prior to the current study measured mainly the recall of information and that other techniques were needed to assess the wide range of abilities included in the definition of competence derived from the critical incident study.

TABLE 6

FACTOR LOADINGS ON THREE ROTATED COMMON FACTORS OBTAINED BY PRINCIPAL COMPONENTS ANALYSIS OF FOURTEEN SUBSCORES ON THE

JANUARY 1966 FINAL CERTIFICATION EXAMINATION

Cumulative Percent of Total Variance	21	32	43
Cumulative Percent of Common Variance	48	75	100

Scores	Communality	I	II	III
Multiple Choice	.55	.70	.25	.09
Short Answer	.43	.65	.13	-.03
Written Simulations				
Problem I Diagnostic Proficiency (Laboratory)	.39	.14	.13	.10
Problem II Treatment Proficiency	.23	.21	.15	.40
Problem III				
Diagnostic Proficiency (Historical Physical)	.46	-.04	.65	-.18
Diagnostic Proficiency (Laboratory)	.63	.13	.79	-.04
Treatment-Proficiency	.12	.20	.08	.27
Conventional Orals				
Pathology	.39	.58	-.14	-.20
Children's	.39	.59	.15	.16
Anatomy and Trauma	.37	.60	.10	.03
Adult	.44	.61	.12	-.21
Simulated Patient Interviews				
Diagnostic Interview Overall	.72	.33	.06	-.78
Proposed Treatment Interview Overall	.69	.42	.01	-.71
Simulated Patient Management Conference Overall	.23	.47	-.11	.00

TABLE 7

FACTOR LOADINGS ON FIVE ROTATED COMMON FACTORS OBTAINED BY PRINCIPAL COMPONENTS ANALYSIS OF 14 SUBSCORES ON THE JANUARY 1968 FINAL ORTHOPAEDIC CERTIFICATION EXAMINATION

Cumulative Percent of Total Variance		17	31	41	52	62
Cumulative Percent of Common Variance		27	50	67	84	100
Scores	Communality	I	II	III	IV	V
Rating Factors						
Information Gathering	.85	.90	.17	.05	.09	.00
Problem Solving	.84	.88	.21	.07	.10	.03
Patient Relationships	.71	.83	-.01	.09	.07	-.06
Multiple Choice						
Recall	.63	.13	.76	.17	-.00	.12
Problem Solving	.54	.10	.72	.10	-.07	.04
Oral Tests						
Trauma-Problem Solving	.44	.09	.56	-.04	.38	-.05
Adult- Problem Solving	.41	.07	.62	-.11	.34	-.11
Child- Problem Solving	.56	.04	.02	.21	.71	-.01
Observation and Interpretation-Interpretation	.41	.19	.17	.41	.36	.24
Simulations-Attitudes	.57	.12	.11	-.09	.73	.01
Written Simulation Exercises						
Diagnostic: Select Indicated Procedures	.77	.07	.28	.65	-.01	-.51
Diagnostic: Avoid Contra-indicated	.73	.02	-.14	-.12	-.11	.83
Treatment: Select Indicated Procedures	.73	.08	-.04	.85	.04	-.00
Treatment: Avoid Contra-indicated	.52	-.04	.29	.09	.13	.64

Analysis of Traditional Oral Examinations

Prior to the current study, certification procedures included an oral examination composed of five half-hour segments, one each in Adult Orthopaedics, Children's Orthopaedics, Trauma, Pathology and Anatomy. Each segment was administered and scored by a team of two examiners. The examiners were responsible for developing questions and supplying any related case materials.

The spirit of the examinations is characterized by the following quotation from the brief set of instructions supplied to all examiners:

> "Try to put a nervous examinee at ease by conversation other than that which pertains to residency and practice. Be fair, do not dwell too long on one subject and cover a variety of materials. It should be apparent very soon whether or not a candidate can answer a question. Try to find out how much he knows."

Given these instructions, it is reasonable to suppose that these examinations were designed to sample breadth of information and to predict that they measured predominantly the recall of information.

Process Analysis

In an effort to make a systematic empirical assessment of the intellectual processes sampled by the orals an observational study, analogous to the process analysis made on the written examination, was conducted on a random sample of the over 2,000 individual oral examinations administered as part of the regular January, 1965 certification procedures.

A team of eight observers (one orthopaedist, two general surgeons, two internists and three professionals in educational evaluation) was trained in systematic observational analysis. For each observation the observer recorded the following information on a form specifically developed for this study (See Appendix 4):

> A verbatim record of each question, the time it was asked, a list of associated visual stimuli (e.g. X-rays, slides), the taxonomic level of the question (i.e., recall, interpretive skill or problem solving), a tally of the number of times the candidate supported his answer (e.g. with an appeal to authority, experience, demonstration, or data), the amount of cueing provided by the examiner, the initial score reported by each examiner and any comments by the examiner or the observer that would clarify the nature of the examination.

Although it is unlikely that the observers were able to record all of the questions asked, 6,868 were recorded in the 158 half-hour observations (including ten duplicate observations), and each was classified according to the intellectual process it seemed to elicit from the candidate. As indicated in Table 8 the degree of inter-observer agreement in the classification of questions was sufficiently high to assure reliable results. These results summarized in Table 9, reveal that:

1. Overall, nearly 70% of the questions appeared to sample only the recall of isolated fragments of information and in one discipline, anatomy, over 90% of the questions were of this type;

2. Fewer than 20% of the questions required the candidate to demonstrate skill in interpreting clinical data (predominantly X-ray).

3. Only 13% of the examiner-candidate exchanges appeared to involve any element of problem solving; and

4. In fewer than 2% of the responses did candidates cite authoritative sources and in only 0.2% of the exchanges did they refer to specific data to support an answer.

It thus appears reasonable to conclude that the traditional oral examinations measure about the same type of competence as the traditional written examinations, and both assess predominantly the ability to recall isolated bits of information.[2]

Statistical Analyses of the Conventional Oral

The results of the process and observational analyses of the traditional examinations led to a decision to subject the orals to extended statistical study during the remaining three years of the project. Statistical data derived from these studies of the reliability and validity of the oral examination are summarized below.[3]

Reliability of the Oral Examination

In principle, there are two major sources of unreliability in the conventional oral : one is due to errors of rating (i.e. different judges will assign different scores to the same performance) and the other is attributable to errors of sampling (i.e. different examiners will pose different questions to the various candidates). An estimate of the first source of error (i.e. interrater reliability) can be obtained by correlating independent scores of two examiners,

TABLE 8
INTER-OBSERVER AGREEMENT, IN PROCESS ANALYSIS OF 1965 ORAL EXAMINATIONS

Observer No.	No. of questions classified as eliciting:			No. of times candidate supported answer by appeal to:	
	Recall	Problem Solving	Interpretive Skill	Authority	Data
(8)	27	10	12	1	0
(1)	31	7	9	2	0
(7)	41	3	12	4	0
(1)	53	3	16	3	0
(8)	53	1	5	0	0
(2)	27	5	5	0	0
(7)	13	7	13	0	0
(2)	1	10	14	0	0
(7)	38	1	17	0	0
(3)	30	11	7	0	0
(8)	13	14	12	2	0
(3)	23	1	8	0	0
(8)	42	13	10	0	0
(3)	50	5	9	0	0
(6)	14	0	13	0	0
(4)	26	3	15	0	0
(5)	30	9	14	3	0
(6)	30	0	11	2	0
(6)	48	4	11	2	0
(7)	43	2	5	1	0

TABLE 9

PERCENT OF QUESTIONS IN 1965 ORAL EXAMINATIONS RATED AT EACH TAXONOMIC LEVEL

	ANATOMY	PATHOLOGY	CHILDREN'S ORTHOPEDICS	TRAUMA	ADULT ORTHOPEDICS	TOTAL
RECALL	94.5	54.8	65.8	66.4	63.4	68.4
PROBLEM SOLVING	4.2	13.6	12.6	17.3	17.1	13.1
INTERPRETIVE SKILLS	1.2	31.6	21.6	16.3	19.5	18.4
NUMBER OF QUESTIONS RATED	1207	1288	1768	1316	1289	6868

both of whom are judging the same performance of a series of candidates; an estimate of the combined effects of the two sources of error (i.e. inter-rater + sampling reliability) can be obtained by correlating the scores of a series of candidates on two different examinations both of which purport to measure the same thing.

In this study an estimate of interrater reliability was obtained by having a team of two examiners administer a single half-hour oral examination in adult orthopaedics to each of thirty selected residents at the time of the 1966 In-Training Examination. The correlation between scores of the two examiners on this series was .72. Under these circumstances pooling of the scores of the two examiners would result in a rating reliability of .90 for the half-hour oral under study. To estimate the combined effects of rating and sampling errors, two half-hour orals in adult orthopaedics were administered to a second sample of 25 residents, by two different examiners. The correlation between scores of the two examiners was .54. Pooling of the scores would yield a coefficient of reliability of .67. Since this figure was obtained from a population that included residents at all levels of training, it is probably somewhat higher than comparable estimates obtained from the more homogeneous population of Board candidates.

As a result of these findings two modifications were made in the administration of the oral examination:

1. Since the number of trained examiners was limited and since the sampling disagreements seemed a greater source of error than rating disagreements, the decision was made to have each examination administered by one examiner (rather than a team of two examiners) in order to maintain or increase the number of independent examinations.

2. The pass-fail decision was to be based on the pooled scores from all oral examinations rather than on scores from each examination considered separately.

Validity of the Oral Examinations

The construct validity of the traditional oral examination was studied by investigation of three hypotheses: (1) Higher scores will be associated with increased education and experience; (2) Assuming that the oral is designed to measure components of competence other than that measured by the multiple choice examination, a factor analysis will show the two loading on different factors; and (3) Assuming that both types of examinations are representative samples of the content specified, correlations between corresponding subject matter sub-tests of the oral and written will be higher than correlations between other sub-tests. Data relevant to the first hypothesis are summarized in Tables 10 and 11. These data were obtained from the 1966 In-Training Examination, in which one conventional oral in Adult Orthopaedics was administered to a selected sample of 233 residents at all levels of training. The results indicate that, as a group, fourth year residents perform substantially better than residents with less training, a finding similar to that obtained on the multiple choice examination (See Section Two, Chapter VIII). However, although mean scores increase with increased amounts of training, there is substantial overlap in performance from year to year. As indicated in Table 11, at least 20% of the second year residents scored higher than the mean score of fourth year residents.

TABLE 10

MEAN SCORES BY LEVEL OF TRAINING
1966 IN-TRAINING EXAMINATION

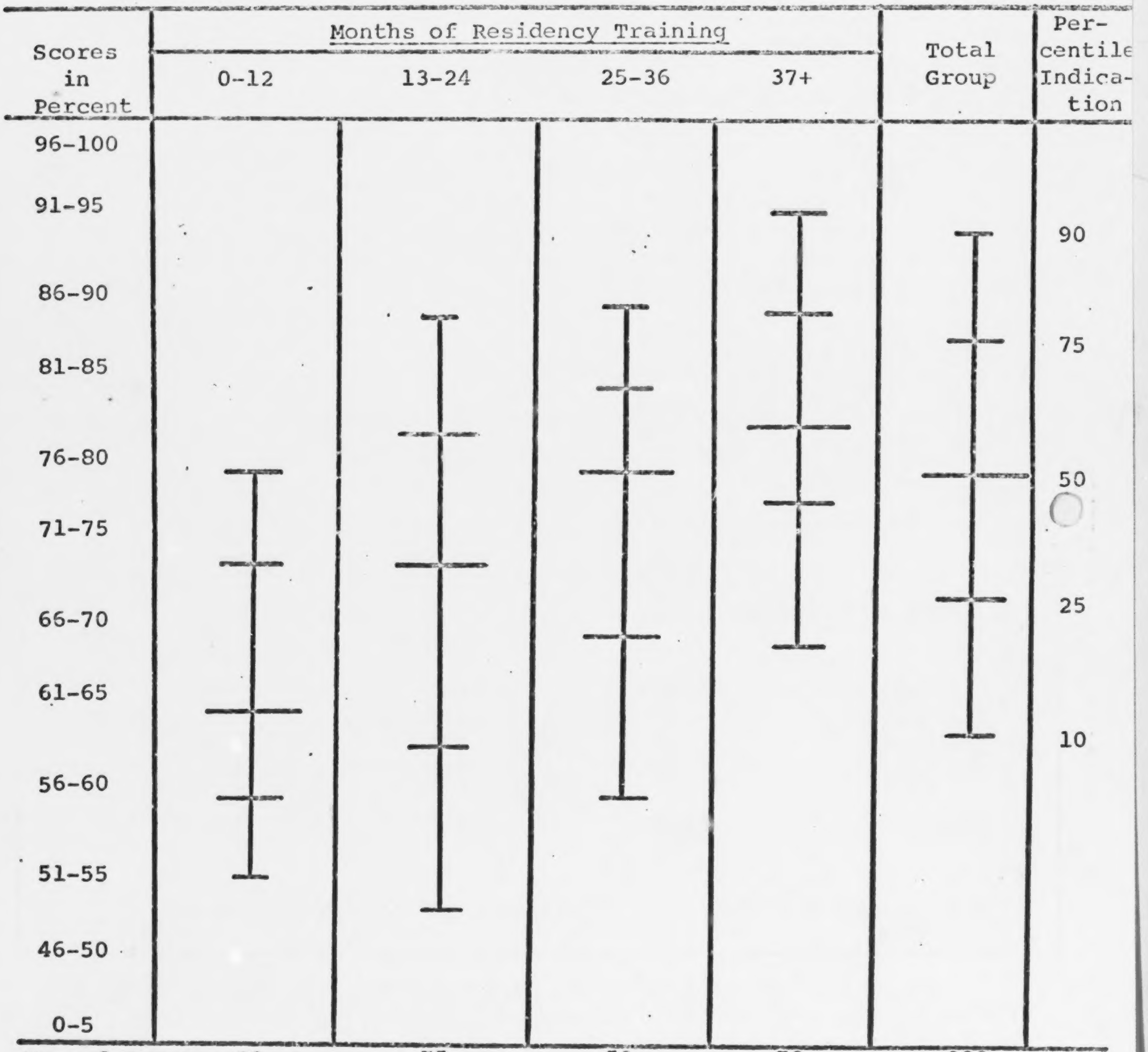
Level of Training	N	Scores	
		Mean *	SD
1st year	29	65%	9%
2nd year	75	70%	13%
3rd year	50	75%	12%
4th year	79	80%	10%
Total	233	74%	12%

* Between group differences significant at .01 level by means of ANOVA

Data on the second hypothesis were obtained from factor analytic studies of the 1966 and 1968 Orthopaedic Certifying Examinations. These studies provided substantial evidence in support of the conclusions of the earlier process analysis of the conventional multiple choice and oral examinations, in that in the 1966 Certifying Examination the two types of examinations showed high loadings on the same factors, whereas, after new types of both written and oral examinations had been introduced

TABLE 11

DISTRIBUTION OF SCORES ON CONVENTIONAL ORAL IN ADULT ORTHOPAEDICS BY LEVEL OF TRAINING*



	0-12	13-24	25-36	37+	Total Group
No. of Cases	29	75	50	79	233
Mean	65.0	70.2	74.6	79.5	73.6

How to read this chart: The column on the left indicates the scores on the examination. The charts indicate the distribution of these scores for each group of residents. The numbers in the percentile column indicate what percentage of the residents got scores below the score on the left. For example, a resident with 14 mos. of experience who received a score of 86 did better than 90% of his fellow residents with equivalent experience.

in the 1968 Certifying Examination the factor structure was considerably more complex and the various components of the examination loaded on somewhat different factors. Further evidence especially relevant to the third hypothesis regarding the inter-relation among sub-test scores was obtained from a study of the 1967 Certification Examination. Since both the conventional multiple choice and the conventional orals were organized on the basis of disciplinary areas it would seem reasonable to assume that the multiple choice subtest in the written should correlate higher with the oral presumed to assess the same content area than with other orals. Table 12 indicates that this assumption is not valid and strongly suggests that the orals are probably measuring a general recall ability independent of any specific content area.

TABLE 12

INTERCORRELATIONS BETWEEN ORAL EXAMINATION SCORES AND
SCORES ON OTHER EVALUATION TECHNIQUES,
1967 ORTHOPAEDIC CERTIFICATION EXAMINATION

N = 351	Multiple Choice		PMP*	Oral Examinations			
Oral Examinations	Total	Adult	Proficiency	Adult	Children's	Trauma	Basic Science
Adult	.33	.27	.09	-	.19	.34	.26
Children's	.41	.29	-.04	.19	-	.37	.33
Trauma	.33	.27	.11	.34	.37	-	.31
Basic Science	.41	.26	.11	.26	.33	.31	-

* Written simulations of Patient Management Problems (PMP), for a description see Section Two, Chapter VI .

Studies of the concurrent validity of various examination techniques, including the conventional oral, were made in connection with the 1966 In-Training Examination which, in addition to conventional multiple choice questions, included a set of simulated problems in patient management (PMP) in the written examination administered to all candidates, and a one hour oral examination administered to 233 selected candidates at all levels of training. This oral examination was divided into three parts: a 30-minute conventional examination in Adult Orthopaedics, a 20-minute simulation of a "diagnostic interview" and a 10-minute simulation of a "proposed treatment" interview with a programmed patient (See Section Two, Chapter VII). Scores on these several components were correlated with training chiefs' ratings of each resident on ten factors representing various aspects of competence. The results shown in Table 13 are by no means easy to interpret, in part, because of the differing reliabilities of the four techniques.

TABLE 13

CORRELATIONS BETWEEN SUPERVISORY RATINGS AND EVALUATION TECHNIQUES, 1966 IN-TRAINING EXAMINATION

Sub-Tests by Technique	First and Second Year Residents (N=107)		Third and Fourth Year Residents (N=119)	
	Rating of Problem Solving	Rating of Overall Competence	Rating of Problem Solving	Rating of Overall Competence
Proposed Treatment Interview	.10	.00	.15	.20
Diagnostic Interview	.14	.12	.23	.16
Adult Oral	.15	.09	.35	.28
Multiple Choice	.23	.20	.26	.26

For example, for third and fourth year residents the Diagnostic Interview which is characterized by very low reliability is almost as good a predictor of ratings of Overall Competence as is the multiple choice sub-test which is characterized by relatively high reliability. Similarly, for this same group, the conventional oral in Adult Orthopaedics is a slightly better predictor of ratings of problem solving ability than is the significantly more reliable multiple choice sub-test. These latter results may be explained in part by the fact that there is heavy emphasis on X-ray interpretation in the oral on Adult Orthopaedics, and X-ray interpretation may play an important role in the training chiefs' definition of problem solving ability.

Results from a subsequent multiple correlational analysis of scores on the 1966 In-Training Examination as predictors of resident ratings by training chiefs are summarized in Tables 14A and B. These data indicate that scores on the conventional oral in Adult Orthopaedics add very little to the prediction of ratings on Factual Information and Overall Competence obtained from the multiple choice examination; in contrast, the two types of tests contribute about equally to the prediction of ratings of problem solving. It is also of interest to note that scores on the oral in Adult Orthopaedics contribute essentially nothing in the prediction of ratings on effectiveness in Patient

TABLE 14

SUMMARY OF RESULTS OF MULTIPLE CORRELATIONAL ANALYSIS USING SUB-TEST SCORES AS INDEPENDENT VARIABLES AND RATING FACTORS AS DEPENDENT VARIABLES

1966 IN-TRAINING EXAMINATION

A. Third and Fourth Year Residents
N = 119

Dependent Variable (Rating Factors)	R	F	Independent Variable (Test Scores)	Partial	
				r	F
Factual Information	.39	2.80**	Multiple Choice Total	.26	7.86**
			PMP, Treatment Problem	-.18	3.50
			Oral, Adult Orthopaedics	.14	2.14
Information Gathering	.53	6.06**	Multiple Choice Total	.37	17.75**
			Oral, Adult Orthopaedics	.22	5.37*
			PMP, Treatment Problem	-.21	5.18*
			PMP, Diagnostic Problem	.19	4.25*
Clinical Judgment	.37	2.47*	Multiple Choice Total	.18	3.72
			Oral, Adult Orthopaedics	.16	2.76
			PMP, Diagnostic Problem	.14	2.07
			Oral, Proposed Treatment Interview	.14	2.05
Surgical Skill	Not Significant				
Patient Relations	Not Significant				
Colleague Relations	.35	2.17*	Multiple Choice Total	.23	6.00*
			PMP, Treatment Problem	-.15	2.40
Ethics	.37	2.51*	Multiple Choice Total	.25	7.31**
			PMP, Treatment Problem	-.18	3.63
			Oral, Adult Orthopaedics	.16	2.89
Overall ompetence	.40	2.94**	Multiple Choice Total	.21	5.21*
			Oral, Adult Orthopaedics	.17	3.22*
			PMP, Treatment Problem	.16	2.91

TABLE 14: (Cont'd)

B. Total Resident Group
N = 228

Dependent Variable (Rating Factors)	R	F	Independent Variables (Test Scores)	Partial r	Partial F
Factual Information	.35	5.19**	Multiple Choice Total	.23	12.80**
			Oral, Adult Orthopaedics	.10	2.29
Problem Solving	.37	7.47**	Oral, Adult Orthopaedics	.18	7.47*
			Multiple Choice Total	.15	5.01*
			Oral, Proposed Treatment Interview	.13	3.52
			PMP, Treatment Problem	-.11	2.82
Information Gathering	.39	6.78**	Multiple Choice Total	.25	14.50**
			PMP, Treatment Problem	.15	5.38*
			Oral, Adult Orthopaedics	.14	4.23*
Clinical Judgment	.34	4.67**	Multiple Choice Total	.17	6.64*
			Oral, Adult Orthopaedics	.13	3.87
			Oral, Proposed Treatment Interview	.10	2.17
Surgical Skill	.25	2.36*	Multiple Choice Total	.12	3.44
			Oral, Adult Orthopaedics	.12	3.32
Patient Relations	Not Significant				
Colleague Relations	.24	2.23*	PMP, Treatment Problem	-.13	3.59
			Oral, Proposed Treatment Interview	.12	2.95
			Multiple Choice Total	.10	2.38
Ethics	.25	2.45*	Multiple Choice Total	.16	6.05*
			PMP, Treatment Problem	-.15	5.18*
Overall Competence	.32	4.16*	Multiple Choice Total	.18	7.26*
			Oral, Proposed Treatment Interview	.10	2.41

* significant at .05 level
** significant at .01 level

Relationships and in Colleague Relationships despite the fact that the oral confrontation could be expected to be indicative of abilities in these areas. In contrast, the new types of orals, especially those that involve simulated interviews with programmed patients, appear to make major contributions in the prediction of these behaviors.

In summary studies of the traditional orals strongly suggested first, that they are slightly, though not significantly more indicative of problem solving and interpretive skills than is the conventional multiple choice examination, and secondly, that the manner in which they are ordinarily used to determine competence entails certain serious, inherent flaws. Revisions of the technique to preserve and enhance its values while minimizing its deficiencies was deemed to be clearly indicated. The new techniques of oral examining developed to meet these needs are described in Section Two, Chapter VII below.

---

1 Benjamin S. Bloom, ed. The Taxonomy of Educational Objective, Handbook I: Cognitive Domain New York David McKay Co. 1956

2 Christine McGuire "The Oral Examination as A Measure of Professional Competence." Journal of Medical Education 41:267-274 March, 1966

3 H. Levine and J. Noak, "The Evaluation of Complex Educational Outcomes" Office of the Superintendent of Public Instruction, State of Illinois, 1968

SECTION TWO

DEVELOPMENT AND ANALYSIS OF NEW EVALUATION TECHNIQUES

CHAPTER IV

STATEMENT OF THE PROBLEM

The critical incident study had indicated that effective behavior as an orthopaedist required the achievement of minimum competence in 94 specific categories of behavior. The analysis of the evaluation techniques used by the Board revealed serious gaps in the procedures used to assess these competencies. The study team was faced with a challenge to devise new techniques or revise old ones in order to assess these competencies, to conduct studies on the validity and reliability of the new techniques, and to assist the Board in incorporating the new techniques in its certification procedures. Furthermore, it was necessary to carry out this research within the framework of the regular examination program conducted by the Board for purposes of certification and by the Academy for purposes of training.

Alternative Approaches to The Development of Assessment Techniques

In initiating the design of such new instruments three possible approaches were considered:

The analytical approach which requires that the elements of the behavior to be measured be carefully defined and means be developed to sample as many of these elements as possible. For example, the ability to recall information about various diseases is one element in the ability to diagnose the causes of a particular constellation of findings; one test of recall will therefore provide information about a necessary but insufficient condition for accomplishing the main objective. Utilizing this approach it is relatively easy to develop exercises in sufficient number to provide a reliable sample of at least one of the specific behavioral elements to be assessed. However, this approach is limited to two respects: First, mastery of the elements of a behavior does not necessarily assure mastery of the total behavior; on the other hand, the elements of a complex behavioral pattern as derived by logical analysis may not be empirically verifiable. For example, diagnosis of a medical problem may actually require much less immediately retrievable information than "logical" analysis would lead one to believe.

The simulation approach to the assessment of complex behaviors involves designing standardized situations that imitiate reality in requiring the examinee to demonstrate the type of behaviors one desires to assess. This approach has the advantage of yielding a direct measure of the complex terminal behavior rather than an assay of its prerequisites or elements and, thus, requires fewer assumptions about the nature of that behavior than does the analytic approach. Further, it is subject to lesser errors of rating and sampling than direct observation of reality. On the other hand, the simulation approach suffers from certain disadvantages: First, to the extent that a "real life" setting differs from a simulation, the simulation may lack validity in predicting behavior in the former situation. Second, because of their complexity, simulation exercises are more difficult to construct and score than simple analytical ones. Third, a simulation exercise requires more testing time per unit than do most conventional analytical exercises and this requires more total testing time to reach a given level of reliability.

The observational approach entails the design of instruments for systematic observation of actual behavior in real situations. This method is especially useful in assessing behaviors which are difficult to simulate or which are not readily amenable to logical analysis into elements. Thus observational techniques are likely to be more valid than other methods of measurement since they obviate the necessity of hypothesizing some relation between the observed behavior and "reality." But these techniques also suffer from certain potential defects: First, the presence of an observer may significantly alter the nature of the behavior being observed; Second, there are inherent problems of sampling and rating observations that make it difficult to achieve reasonable levels of reliability with such methods; and finally, certain important aspects of behavior are quite difficult to observe, even in "reality."

Further, it should be noted that the observation of behavior is of two types: One, observation and recording of a specific sample of behavior (e.g. a patient interview) and, two, the observation and rating of the examinee's habitual behavior. These two types have different advantages and disadvantages: In the first type it is often possible to minimize examiner bias by using standardized observational techniques and by training observers; however, since situational variables are so important in some behaviors, it may be quite impossible to generalize about an individual's competence from only one or two specific observations, furthermore, the potential distortions attributable to the presence of an observer are maximized in this type of test situation. In contrast, observation and rating of habitual performance suffers greatly from observer bias, and some persons who are in an ideal position to know about habitual performance are not capable of rating it objectively; however, when observations are made by skilled raters in a position to observe an examinee's behavior over long periods of time, they have maximum generalizability and validity. Finally, it is obvious that certain aspects of performance (e.g. ethical attitudes) are not readily susceptible to other modes of evaluation.

Specific Instruments Developed

In the Orthopaedic Training Study, all of the approaches described above have been employed in designing the new types of evaluation instruments listed in Table 15. Once a new method was devised, experimental instruments were developed and administered to various populations in order to obtain evidence regarding the validity and reliability of the

TABLE 15

LIST OF TECHNIQUES UTILIZED DURING THE ORTHOPAEDIC TRAINING STUDY

TECHNIQUE	COMMENTS
I. Analytical A. Multiple Choice Tests	These exercises require the examinee to demonstrate problem solving and interpretive abilities as well as recall of information.
II. Simulation A. Written Simulations	These exercises require the examinee to demonstrate skill in solving diagnostic and/or treatment problems involving sequential analysis and decision; they employ a special answer sheet with an erasable overlay designed to provide feedback to the examinee about the results of his inquiries and therapeutic interventions.
B. Oral Tests of Complex Cognitive Behavior	The 5 types of exercises, described below, though not strictly simulations, resemble them in many ways and share the same characteristics.
1. Diagnostic Problem	This type of exercise requires the examinee to arrive at and defend a diagnosis on the basis of information he elicits by inquiries made to an oral examiner.
2. Defense of Therapy Problem	This type of exercise requires the examinee to present his rationale for the therapeutic decisions he makes on a standardized case presented to him.
3. Emergency Treatment Problem	This type of exercise requires the examinee to describe the procedures he would follow in treating a specific emergency case in the emergency room, and the actions he would take in responding to the consequences of his therapeutic interventions and diagnostic inquiries as reported by the examiner.

4. Complication Problem	This type of exercise is similar to the emergency treatment problem, except that it deals with problems of long term care.
5. Observation and Interpretation Exercise	This type of exercise requires the examiner to describe what he sees in slides and x-rays and to relate these to other data about a specific case.
C. Oral Simulations of Interpersonal Conferences	This type of exercise requires the examinee to play the role of a physician while the examiner assumes the role of a patient, colleague, or paramedical person in typical interpersonal confrontations in the practice of medicine; the exercises are designed to evaluate the candidates'ability to communicate with and relate effectively to patients, colleagues and paramedical personnel.
D. Oral Simulations of Group Conferences	This exercise requires 5 examinees to simulate a staff Conference on the management of 2 specific cases.
OBSERVATIONS A. Habitual Performance 1. Candidate Evaluation Form	This form is designed to obtain supervisor's rating of various aspects of the habitual performance of candidates for Board certification.
2. Resident Evaluation Form	This form is similar to the Candidate Evaluation Form, but is designed to obtain supervisor's ratings of resident performance.
B. Samples of Performance 1. Rating Form for Assessing Surgical Skills	This form is a detailed checklist to be used in the observation of any surgical procedure.
2. Rating Forms for Evaluating Behavior in Oral Examinations	These forms specify and define the criteria to be employed in rating various aspects of performance on each type of oral examination.

technique. As noted earlier most of these reliability and validity studies were carried out in connection either with the final Certification Examination administered by the American Board of Orthopaedic Surgery for purposes of specialty board certification to all candidates who have completed an orthopaedic residency and one year of practice, or in connection with the In-Training Examination administered by the American Academy of Orthopaedic Surgeons for diagnostic and feedback purposes to virtually all residents in the United States training programs. The specific examinations on which studies were conducted are listed in Table 16.

Data Analysis

Studies of the reliability of the newer objective techniques (i.e., the multiple choice and written simulation exercises) were conducted by employing adaptations of conventional methods for estimating internal consistency. Studies of the interrater reliability of techniques involving subjective judgments, (i.e., orals and rating forms), were conducted by obtaining and correlating independent ratings of two observers judging the same behavioral sample; estimates of the combined effects of rating and sampling errors were obtained by correlating the independent ratings of two observers judging different behavioral samples.

Data on the content validity of the several techniques were collected by process analysis. Data on construct validity were obtained by studying the relationship between test scores and such examinee characteristics as age, practice setting, and amount of training utilizing analysis of variance methods. Additional data on construct validity were obtained from correlational and factor analytic studies designed to test hypotheses regarding interrelationships among scores on the various types of exercises. Data on concurrent validity were obtained by multiple correlational analysis of the relationships between ratings of habitual performance and scores on the several types of exercises.

Though many techniques were, of necessity, analyzed simultaneously, the data on each will be presented separately and, in the interests of clarity, will be reported throughout this section in the following order:

1. A description of the technique, a brief history of its development and some basic considerations in construction of exercises of the type described;

2. Data and discussion of the reliability of the technique;

3. Data and discussion of the content, construct and concurrent validity of the technique.

TABLE 16

EXAMINATIONS STUDIED DURING ORTHOPAEDIC TRAINING STUDY

EXAMINATION	POPULATION	COMMENTS
(1) May 1965 Certification	406 Candidates	First extensive study of written simulation exercises.
(2) Nov. 1965 In-Training	1,398 Residents	First study of In-Training Examination
(3) Jan. 1966 Certification Final	461 Candidates	First study of new oral examinations and continuation of study of written simulation exercises.
	184 Examiners	Study of construct validity of the written simulations administered to candidates
(4) May 1966 Certification I	459 Candidates	First attempt to obtain ratings of habitual performance from training chiefs in a study of written simulation exercises and multiple choice questions.
(5) Nov. 1966 In-Training	1,539 Residents	Validation of simulation exercises and multiple choice questions by use of ratings of habitual performance and cross-sectional analysis of performance.of sub-groups.
	233 Candidates	Orals administered to sample of residents to obtain reliability and validity data on all techniques.
(6) Jan. 1967 Certification Final	449 Residents	Replication of Jan. 1966 study of the Final Certification examination.
(7) Nov. 1967 In-Training	1,682 Residents	Replication of Nov. 1966 study of the In-Training Examinations.
(8) Jan. 1968 Final Certification	838 Candidates	First complete implementation of revised certification procedures including the adoption of a profile system for reporting scores; reliability and validity studies of all techniques.
TOTAL	7,480	Examinees who participated in more than one study are tallied separately in each.

The two final chapters of this Section contain, first, a report of the data and discussion of the interrelationships among all techniques, and second, a description of the system developed by the American Board of Orthopaedic Surgery for incorporating all techniques in its current certification procedures.

CHAPTER V

OBSERVATIONAL FORMS

As noted earlier (See Table 15), most of the new types of examination exercises developed in the Orthopaedic Training Study represent efforts to simulate the important behaviors described in the critical incident study. Since simulations cannot be perfect copies of reality, it is necessary to demonstrate that the correspondence between behavior in actual situations and in the simulated situations is such as to justify using the latter--for purposes of estimating competence in the former. For this reason, observational techniques have been extensively employed as criterion measures in this study, despite the deficiencies in reliability and validity to which they are often subject. However, to the extent that data from observations agree with data from simulations, increased confidence can be placed in both.

Forms for Rating Habitual Performance

Ratings of habitual performance have been utilized in this study for three major purposes: (1) to provide data on the concurrent validity of other types of evaluation techniques; (2) to provide data on aspects of competence not assessed by other techniques; and (3) to educate training chiefs regarding the dimensions and complexity of competence in orthopaedic surgery and to assist them in monitoring the progress of residents in achieving these goals.

For these purposes two different observational forms were developed. The Resident Evaluation Form (Appendix 5) was developed primarily as part of the In-Training Evaluation of residents. It was designed to obtain evidence on the following factors: ability to recall factual information concerning general medicine and orthopaedic surgery, ability to use information to solve problems, ability to gather clinical information, judgment in deciding on appropriate treatment and care, skill in surgical procedures, relating effectively to patients, relating effectively to colleagues and other medical personnel, demonstrating the moral and ethical standards required of a physician, and overall competence as an orthopaedic surgeon. A revised version of this form has been incorporated as part of the In-Training Examination and is completed each year by the resident's chief of training (Appendix 6). The Candidate Evaluation Form (Appendix 7) was developed for rating of candidates for certification; it is an adaptation of the earlier Resident Rating Form designed to yield evidence on the following factors: information gathering, problem-solving, clinical judgment, surgical technique, relating to patients, continuing responsibility, emergency care, relating to colleagues, moral and ethical values, and overall competence.

These two forms represent somewhat different ways of resolving the problems of obtaining valid and reliable ratings.

The first of these problems is that of specifying and describing the factors on which ratings are to be made so as to avoid having behavior of the same type classified differently by two raters. The two sample descriptions of factors reported in Table 17 illustrate the alternative ways in which this problem was resolved in the two forms under discussion.

The second problem concerns the nature of the scale to be used in recording the ratings. You will note that a 12-point scale was utilized in both forms in order to faciliate pooling of data from a variety of techniques for purposes of making an overall judgment of a candidate's competence.* However, once that scale had been decided on different methods were employed in the two forms to try to maximize agreements between raters regarding the meaning of each point on the scale. The reader will note (Table 17) that in the Candidate Evaluation Form this problem was dealt with by describing the extreme points of the scale in behavioral terms and by placing adjectival meanings at intermediate points. While this is a defensible procedure for ratings in a certification examination, it creates certain difficulties because raters are often reluctant to use the negative half of the scale. The consequent restriction of range lowers the reliability of the ratings. For example, among 1574 ratings on the "Overall Competence" of candidates applying for the 1968 Certifying Examination only 5 were in the "poor" range of the scale, only 78 in the "marginal" range, but there were 174 additional ratings at "7" which is the point on the scale just above "marginal".**

Because of this "error of leniency" the adjectives were eliminated from the Resident Evaluation Form and the raters were instead instructed to rank residents as follows:

> "In filling out this form you are to rank the resident on each factor in terms of all the residents in orthopaedic surgery you have known during your career. You are to indicate your rankings by checking the appropriate box under each factor. In making these evaluations DO NOT take into account the resident's level of training. For example, a second year resident may have the potentiality to display outstanding surgical skills, but many fourth year residents might function AT THE PRESENT time on a higher level. He should be ranked lower than they are ranked on surgical skill."

---

* Data from oral examinations and from observations are collected directly on the 12-point scale. Data from written examinations (i.e., multiple choice and simulation exercises) are converted to the 12-point scale; for a description of the conversion technique see Section Two Chapter X .

** See Appendix B for complete set of rating data on the 1968 Certifying Examination.

TABLE 17

DESCRIPTION OF THE FACTOR: JUDGMENT IN TWO RATING FORMS

A. Resident Evaluation Form

- -

Col. No.

Factor IV: Judgment in deciding on appropriate treatment and care

This factor deals with the resident's ability to properly weigh the many factors involved in deciding on treatment and care, and to come to sound conclusions.

19 I do not have sufficient information to judge. 1 ☐

RANKING

20-21

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

TABLE 17 (con't)

DESCRIPTION OF THE FACTOR: JUDGMENT IN TWO RATING FORMS

B. Candidate Evaluation Form

Col. No.	Factor 3. CLINICAL JUDGMENT	
	This factor is concerned with the Candidate's ability to use sound judgment in planning for and carrying out treatment.	
	The INEFFECTIVE Candidate is overly concerned with treatment techniques at the expense of overall goals.	The EFFECTIVE Candidate is familiar with the uses and limitations of the procedures he attempts. He recognizes his own capabilities and uses procedures which correspond to them.
	He often delegates pre- and post-operative care to others.	He considers simple procedures first.
	He plans treatment without sufficient familiarity with the procedures he selects.	His clinical judgment encompasses information beyond the pathologic.
	His treatment choice is rigid--using a set formula for treating each clinical problem or using a favorite technique when more effective ones are available.	He demonstrates regard for patients' needs, desires and life conditions.
		He is flexible enough to modify his treatment plans when the situation warrants doing so.

Col. No.												
48-49	01	02	03	04	05	06	07	08	09	10	11	12
	Poor			Marginal			Good			Excellent		
50	☐ Insufficient information to judge											

While this technique tends to reduce the "error of leniency" it suffers from the defect that each rater must employ standards based upon the sample of residents he has met, and these samples will differ from program to program. Where each judge rates a relatively large group, errors arising from this source can sometimes be minimized by normalizing the distribution of ratings from each observer and converting all to standard scores; this approach was not possible in this study due to the fact that each supervisor rated very few cases. For this reason every effort was made to obtain multiple ratings on each individual; this too proved to be unfeasible in this study. The observational data presented below are therefore based on the pooled ratings of only two supervisors.

Reliability of Observational Forms

In studying the reliability of the Resident Evaluation Form ratings of each resident were obtained from two supervisors in the same program; whereas in studying the reliability of the Candidate Evaluation Form, in general, the ratings of each candidate were obtained from training chiefs in different programs; these data could therefore be expected to have maximum generalizability.

The results of these studies, as summarized in Table 18, indicate that the Candidate Evaluation Form is much less reliable than the Resident Evaluation Form. This finding can be attributed to several factors. First, a candidate probably does behave somewhat differently in different programs; it is therefore reasonable to expect a lower correlation between ratings in different programs than between those obtained in the same program. Second, raters are more likely to agree on standards within programs than across programs. Third, raters within a program often discuss resident performance and thus influence each others ratings; this is far less likely to happen across programs. Fourth, residents do, in fact, differ more than candidates simply because of the greater variation among them with respect to education and experience; this increased range of competence will in itself tend to increase the reliability of the ratings. Finally, supervisors are less hesitant about giving low ratings to residents than to candidates and this, too, will increase the range, and thus increase the reliability of resident ratings as compared with candidate ratings.

Given these considerations it seems reasonable to conclude that the reliability of the ratings can be significantly raised only if there is opportunity to increase the number of ratings per candidate and to institute a training program for raters. It is therefore of interest to note that in incorporating the rating forms in the regular certification procedures, the Board is doing so in a manner designed to produce these improvements in the collection of observational data.

TABLE 18

RELIABILITY DATA ON RATING FORMS

A. Resident Evaluation Form (N=190)

FACTOR	MEAN OF BOTH RATINGS	SD OF BOTH RATINGS COMBINED	CORRELATION BETWEEN RATERS	RELIABILITY*
1. Recall of Factual Information	8.0	1.9	.33	.50
2. Problem-Solving	8.3	1.8	.53	.69
3. Information Gathering	8.4	1.7	.45	.62
4. Clinical Judgement	8.1	2.0	.53	.69
5. Surgical Skill	8.3	1.8	.35	.52
6. Patient Relationships	8.8	1.8	.37	.54
7. Colleague Relationships	8.8	2.0	.50	.67
8. Ethics	10.2	1.5	.49	.67
9. Overall Competence	8.2	2.0	.56	.72

B. Candidate Evaluation Form (N=391)

FACTOR	MEAN OF BOTH RATINGS	SD OF BOTH RATINGS COMBINED	CORRELATION BETWEEN RATERS	RELIABILITY
1. Information Gathering	9.1	1.4	.17	.29
2. Problem Solving	9.0	1.5	.17	.29
3. Clinical Judgement	9.1	1.4	.19	.22
4. Surgical Technique	9.3	1.3	.17	.29
5. Patient Relationships	9.3	1.5	.14	.25
6. Continuing Responsibility	9.5	1.4	.13	.23
7. Emergency Care	9.6	1.3	.16	.28
8. Colleague Relationships	9.4	1.5	.17	.29
9. Moral and Ethical values	10.2	1.4	.15	.26
10. Overall Competence	9.3	1.3	.18	.31

* Computed by the Spearman-Brown Formula.

Validity of Observational Forms

Content Validity.

Since the factors to be rated in the Candidate and Resident Evaluation Forms were derived directly from the specific components of competence identified in the Critical Incident Study and refer to habitual, observable performance, these forms are, by definition, characterized by high content validity. In short, in using rating forms of the type described above, it is unnecessary to make any assumption about the relation between the behaviors sampled by the instrument and those demonstrated in "real life" situations.

Despite this obvious advantage, they, like most other rating forms, are subject to certain deficiencies which can reduce their validity significantly. One such defect derives from the tendency of some raters to rate individuals who are high or low in one trait as high or low in all traits (the halo effect). Of special importance in this regard is the tendency to credit the person with a pleasing personality with greater cognitive skills than he has actually achieved. Since an important aspect of competence for physicians and other professionals who must constantly deal with people is the ability to impress others with their competence and dedication, this specific halo effect may not be as serious as it seems; it nevertheless reduces the validity of the rating to the extent that it results in inflated evaluations of purely cognitive or psychomotor skills. Second, in using the forms under discussion, raters are sometimes guilty of logical errors in assuming a closer relation between certain attributes, for example, problem-solving skills and clinical judgment, than is warranted. High ability in one such factor may lead to undeservedly high ratings on the other. Third, raters sometimes rate examinees on attributes of behavior which they have not directly observed. In such cases they tend to arrive at a judgment on the basis either of a general "halo" or the type of logical error discussed above. For example, in this study, statistical analysis of the ratings leads to the suspicion that in rating "information gathering ability" supervisors were actually rating "diagnostic ability." This suspicion is supported by the fact that residents are rarely observed gathering information, i.e. interviewing a patient, for example. Finally, ratings are sometimes affected as much by the inadequacy of raters as by the ability of the examinees; some raters are systematically too lenient, others too harsh, and others indiscriminating.

For the reasons outlined above ratings must be considered as simply another evaluative technique and not as the ultimate criteria against which all other evaluation techniques must be judged. However, since ratings are obtained in a fashion so different from other evaluative techniques, agreement in the results from two such sources supports the view that there must be some underlying behavioral manifestations which account for the congruence and which thus help to confirm the validity of both the ratings and the other evaluation techniques. It is from this point of view that studies of concurrent validity have been conducted using ratings as criteria.

Construct Validity

Two types of studies were conducted to investigate the construct validity of ratings of habitual performance. The first entailed analysis of the relation between level of training and level of ratings. This study revealed slight differences in the expected direction. For example, utilizing a 12-point scale the mean rating on "Overall Competence" was 7.7 for residents with 1-2 years' training and 8.1 for residents with 3-4 years' training. While this difference is statistically significant for the number of cases included in the study, one must question its practical significance.

The second study of construct validity of the ratings entailed analysis of the interrelationship among scores on the several factors, to determine the amount of halo in each. Table 19 summarizes the important correlational data for both the Resident and Candidate Evaluation Forms. It indicates some independence of factor scores but also reveals a strong halo effect. This independence is more marked in the Resident Evaluation Form than in the Candidate Evaluation Form, probably due to higher reliability of the former. Note, for example, that the correlation between ratings of Surgical Technique and Patient Relationships on the Resident Form ranges between .35-.43 when the same rater judges both, but only .17 across raters; the correlation between ratings on these two factors on the Candidate Evaluation Form is .58-.62 for the same rater and .10-.12 across raters. Further, it is important to note that as a consequence of both the halo effects and the differential reliabilities of different factor ratings, the correlation between two ratings of the same factors is in some instances lower than the correlation between two judges' ratings of different factors. For example, on the Resident Evaluation Form the correlation between two ratings of Surgical Technique is .35 while the correlation between one rater's rating of Surgical Technique and a second rater's rating of Problem Solving is .33-.34. This type of correlational pattern is even more pronounced in the Candidate Evaluation Form.

In view of these findings it is of particular interest to examine the relationship between ratings on various factors of habitual performance and scores on tests designed to measure these same behavioral factors. Table 20 presents illustrative data of this type.* It indicates, as would be expected, that scores on the recall component of the multiple choice examination are less closely associated with ratings of "Patient Relationships" than with ratings of "Problem-Solving" skill. Much of the data of this type on the concurrent validity of the rating forms is based on the assumption that the forms are valid, and that the validity

---

* For a full discussion of this aspect of concurrent validity of all techniques, see Section Two Chapters VI through VIII.

TABLE 19

RELATIONSHIPS BETWEEN SELECTED RATING FACTORS IN BOTH CANDIDATE AND RESIDENT EVALUATION FORMS

Resident Evaluation Form N=190

	FACTORS							
FACTORS RATERS	PROBLEM SOLVING I	PROBLEM SOLVING II	SURGICAL TECHNIQUES I	SURGICAL TECHNIQUES II	PATIENT RELATIONSHIPS I	PATIENT RELATIONSHIPS II	OVERALL COMP. I	OVERALL COMP. II
PROBLEM SOLVING I	-	.53	.52	.33	.44	.28	.75	.50
PROBLEM SOLVING II	.53*	-	.34	.66	.28	.42	.50	.75
SURG. TECH. I	.52	.34	-	.35*	.35	.17	.55	.30
SURG. TECH. II	.33	.66	.35*	-	.17	.43	.29	.57
PATIENT REL. I	.44	.27	.35	.17	-	.37*	.47	.35
PATIENT REL. II	.28	.42	.17	.43	.37*	-	.28	.41
OVERALL COMP. I	.73	.44	.55	.29	.47	.28	-	.56*
OVERALL COMP. II	.50	.75	.30	.57	.35	.41	.56*	-

Candidate Evaluation Form N=391

FACTORS RATERS	PROBLEM SOLVING I	PROBLEM SOLVING II	SURGICAL TECHNIQUES I	SURGICAL TECHNIQUES II	PATIENT RELATIONSHIPS I	PATIENT RELATIONSHIPS II	OVERALL COMP. I	OVERALL COMP. II
PROBLEM SOLVING I	-	.17*	.73	.16	.63	.11	.85	.17
PROBLEM SOLVING II	.17*	-	.16	.67	.11	.62	.17	.81
SURG. TECH. I	.73	.16	-	.17*	.62	.10	.80	.18
SURG. TECH. II	.16	.67	.17*	-	.12	.58	.14	.75
PATIENT REL. I	.63	.11	.62	.12	-	.14*	.77	.18
PATIENT REL. II	.11	.62	.10	.58	.14*	-	.11	.71
OVERALL COMP. I	.85	.17	.80	.14	.77	.11	-	.18*
OVERALL COMP. II	.71	.81	.18	.75	.18	.71	.18*	-

* Correlation between two ratings of the same factor across raters represents an estimate of the reliability of one rating.

of the testing techniques can be assessed by analyzing their value in predicting ratings, treating the latter as dependent variables. Such analyses therefore properly belong with the discussions of the individual techniques that are treated as independent variables. However, given the factorial structure of certain evaluation techniques it is also appropriate to analyze the pattern of interrelationships ratings and test scores where the various rating factors are treated as the independent variables.

TABLE 20

ILLUSTRATIVE CORRELATIONS BETWEEN SELECTED RATING FACTORS and TEST SCORES

1968 ORTHOPAEDIC CERTIFICATION EXAMINATION
N=391

Rating Factors	Test Scores Multiple Choice Recall	 Oral Mean Problem Solving	 Simulation Attitudes
Problem Solving	.33	.31	.19
Surgical Technique	.16	.21	.10
Patient Relationships	.10	.21	.15

Such data are summarized in the multiple correlational analysis reported in Table 21. These data reveal that the rating form factors have different weights when used to predict different test scores. It is especially interesting to note the effects of variables that have negative partial r's. For example, note that Overall Competence and Problem-Solving account for most of the correlation between the Multiple Choice Recall score and the rating factors, and that this correlation is improved when scores on Patient Relationships and Surgical Technique are given negative weights. This phenomenon is probably explained in part by the fact that in rating Overall Competence the supervisor takes into account the cognitive, psychomotor and affective skills of the subject; if he perceives two men as equally competent, the man with higher affective and psychomotor skills will be likely to have lower cognitive skills since the chief's perception is the result of an amalgam of all three components of competence.

The multiple R's reported in Table 21 and in the subsequent tables in this section may seem dismayingly small, and such a conclusion would be justified if the tests had been developed for purposes of predicting supervisors' ratings. Rather they are intended as measures of competence for purposes of certification. The studies of concurrent validity

TABLE 21

RESULTS OF MULTIPLE CORRELATION ANALYSIS OF
JANUARY 1968 CERTIFICATION EXAMINATION USING RATING FACTORS
AS INDEPENDENT VARIABLES AND VARIOUS TEST SCORES AS DEPENDENT VARIABLES

DEPENDENT VARIABLES : TEST SCORES	R	F	INDEPENDENT VARIABLES: RATING FORM FACTORS	PARTIAL a	F	SIMPLE r
Multiple Choice, Recall subscore	.34	5.42**	Overall Competence	.14	7.39**	.26
			Problem Solving	.12	5.20**	.28
			Patient Relationships	-.10	3.68	.10
			Surgical Technique	-.09	3.32	.13
Oral Exam in Trauma, Problem Solving subscore	.26	3.10**	Information Gathering	.10	3.46	.21
			Continuing Responsibility	-.09	3.17	.11
Oral Exam in Observa-Interp; Interp Subscore	.28	3.57**	Problem Solving	.13	6.16*	.27
Written Simulation of Diagnotic Type, Subscore Selection of Indicated Procedures	.22	2.05*	Surgical Technique	-.10	4.07*	.06
			Patient Relationships	.08	2.21	.18

* Sig. at .05 level
** Sig. at .01 level

NOTE: Data are not given for Independent Variables with F-Ratios below 2.00.

have therefore been designed to test the hypothesis that the tests measure important areas of competence. Second in interpreting the correlations between rating factors and test scores it is necessary to recognize that the reliabilities of sub-scores of each are in some cases quite low. For example, the reliability of the multiple choice recall score is .71 and in no case does the reliability of a rating factor exceed .31. (See Table 18) Given this much error in the two sets of scores a multiple correlation of .34 is of considerable practical, as well as theoretical, significance.

In summary, the rating form data in the present study have been useful primarily as a means of gathering evidence on the validity and the factor composition of the various test techniques. The results to date suggest that by increasing the number of observations, and by specifying the factors to be rated in more operational detail, the reliability of the ratings can be raised sufficiently to justify their use in evaluating the relative effectiveness of varied curricular settings. Ultimately, however, the main impact of the observational techniques developed for this study will depend on the degree to which they are of assistance to those responsible for training in defining program objectives in more specific terms, in monitoring residents' progress and in diagnosing the strengths and weaknesses of both the resident and the program.

Ratings of Specific Incidents of Performance

Observational Rating of On-the-Job Performance

In addition to ratings of relevant aspects of habitual performance it is often useful to record an examiner's evaluation of a specific incident of on-the-job performance. In the current study only one such rating form, The Form for Evaluation of Surgical Skill, has been developed. An excerpt from that form* is shown below, illustrating the specification, the description, and the method of rating one factor.

Factor I. Initial preparation for surgery

Did surgeon reaffirm procedure with patient before surgery? Was surgeon properly gowned? (Cap over hair, mask tight, careful timed scrub of hands, gets into gown and gloves properly.) Did he review procedure and position with anesthesia staff? Was the position of patient appropriate for procedure. (Surgery, tourniquet, X-rays, bone graft, etc.) Did he know the names of nursing and medical staff? Was preparation of patient's skin adequate? (Check area scrubbed, technique of applications, method of discarding sponges, etc.) Was draping satisfactory and appropriate to procedures? Did he prevent contamination by others?

4 ☐ Excellent Comments
3 ☐ Good
2 ☐ Adequate
1 ☐ Poor

Note, that there are alternative possible ways of scoring performance: For example, it would be possible to require the observer simply to answer "yes" or "no" to each of the questions listed in the excerpt. Such a method could be expected to achieve high reliability in initial ratings; however, it would still be necessary to determine (presumably, on the basis of the "yes" and "no" answers) whether the surgeon had mastered the procedure at a satisfactory level. In the present study the decision was therefore made to utilize the questions merely to define the factor to be evaluated and to require the examiner to make a direct value judgment about the adequacy of the behavior observed.

To date, this form has been employed only experimentally in small pilot studies; no statistical data are as yet available on the reliability and validity of ratings derived from it.

---

* See Appendix 9 for the complete text of The Form for Evaluation of Surgical Skill.

Observations of Oral Simulations and Oral Tests of Complex Cognitive Behavior

Since oral simulations are complex performances, the problems of developing and applying rating forms for scoring oral examinations are similar to those involved in evaluating on-the-job performance.

The effectiveness of such examinations depends upon two conditions: First, the ability of the examiners to develop techniques which elicit behaviors that do, in fact, sample important areas of competence; and second, the development of rating procedures that will yield reliable assessments of the behaviors elicited by the techniques. The first condition is discussed in subsequent chapters in connection with summaries of the studies of oral examinations; the second condition is the main issue of concern here.

During the course of the Orthopaedic Training Study, four types of forms for rating oral examinations were developed.* Two approaches are represented in these forms. The first approach is illustrated by the following excerpt taken from the "Rating Form for Use with Patient Interviews." The reader will note that this approach is similar to that discussed above in the rating of habitual performance, and like its counterpart, can suffer from various types of errors of classification, e.g., one examiner's "04" is equivalent to another's "07" even though both agree on what the examinee did. However, it is important to observe that so long as this represents systematic differences in standards, errors of this type do not influence the correlation between the scores of two raters.

Factor I: Ability to elicit an adequate amount of pertinent information

(The candidate should ask most of the indicated questions; other questions should be appropriate to the diagnosis.)

01	02	03	04	05	06	07	08	09	10	11	12
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	Poor			Adequate			Good			Excellent	

In the present study it was found that in utilizing the form illustrated above, the correlation between two examiners' ratings of the same set of half-hour oral examinations was generally about .70.** It would therefore appear that much of the difficulty in generalizing from oral examination data is due not so much to errors of perception in human judges, but to errors of classification and sampling. This is an important finding since classification errors can be reduced by statistical correction and by pooling data from many examiners, while sampling errors can be reduced by enlarging the sample of examinee behavior observed. Alternatively, if the judges disagree seriously in the ranking of examinees, it is impossible to use the ratings as measures of examinee competence, since they will reflect rater bias rather than the "real" behavior of the examinee.

---

* See Appendices 10 through 13 for copies of these forms.

** For a detailed summary of the data see Section Two, Chapter VII.

RATING SCALE

INSTRUCTIONS: Place a tally mark in the appropriate box for EACH statement EACH candidate makes.

LEVEL OF STATEMENT	CANDIDATE NO:				
	A	B	C	D	E
O. ERROR					
I. HINDRANCE					
II. PASSIVITY					
III. CLARIFICATION					
IV. INTEGRATION FACILITATION					

The second approach to the development of rating forms is illustrated in the excerpt from the "Rating Form for Simulated Patient Management Conferences" quoted above.* In using this form the examiner is directed to record the number of times that a specified behavior occurs. On theoretical grounds it would appear that this approach would tend to maximize inter-observer reliability despite the fact that observers will disagree on the classification of some specific behavioral incidents. Unfortunately, it was not possible in the present study to obtain data on the use of this method with a sufficient number of cases to test this hypothesis.

Summary Comment

The objectified forms developed in this study for rating both habitual performance and oral examinations appear to provide valuable information on aspects of examinee performance not readily assayed by written examination. However, both types of ratings are subject to errors of classification and sampling which must be taken into account in generalizing the data derived by such methods. With due allowance for such errors, the forms developed for use in this study yield ratings of resident performance at different levels of training that differ in the expected direction, and ratings of candidate performance that are associated in the hypothesized ways with estimates derived from more objective written exercises.

---

* For a description of this examination and the complete text of the form see Section Two, Chapter VII, and Appendix 12.

CHAPTER VI

WRITTEN SIMULATION EXERCISES

Statement of the Problem

In the critical incident study a number of activities relating to the processing of information were identified as essential ingredients of competency in the practice of orthopaedic surgery. The following are illustrative of these critical requirements:

1. Obtaining adequate information from the patient
2. Consulting other physicians
3. Checking other sources
4. Directing or ordering appropriate films
5. Obtaining biopsy specimen
6. Persisting to establish definitive diagnosis

The study further identified some of the elements involved in the ability to make appropriate decisions in complex situations. The following are illustrative:

1. Indicating suitable treatment for condition
2. Treating with regard to special needs
3. Choosing wisely between simple and radical approach
4. Delaying therapy until diagnosis better established
5. Treating most critical needs first
6. Reassessing, altering or repeating treatment

However, both the process and the statistical analyses of the commonly used oral and written techniques revealed that they did not yield adequate assessments of most of these skills and abilities. In the clinical setting, the individual problem-solver is usually confronted initially with only very limited information, such as the presenting complaint of a patient. From this information he must generate a hypothesis, gather data and, on the basis of these data generate new hypotheses. Sometimes the most important component of competence consists in the ability to carry out the process of hypothesis testing effectively. Less competent individuals may fail either by coming to premature conclusions or by refusing to make a decision when the situation demands that type of behavior. Furthermore, they may misinterpret results, pursue false hypotheses with stubborn persistence or make serious errors in judgment about the significant data they need to obtain or about the findings they do collect or about the relative weight of relevant factors in arriving at a decision. In sharp contras with this reality most conventional written examinations provide significant cues to the examinee by presenting him with a limited amount of information which is, by definition, adequate for the solution of the problem posed--a form of cueing rarely present in real life. To avoid this distortion of reality various simulation techniques have been devised for assessment of certain aspects of professional competence relating to clinical judgment in orthopaedic surgery.

Description of Written Simulations

The written simulations developed in the current study employ a special answer sheet with an erasable overlay which can be used to give immediate feedback to examinees.[1] The problems are then designed to require the examinee to make choices from an almost unlimited number of broad strategic routes, several of which may lead to an acceptable result.[2]

A typical problem is initiated by a brief description in either verbal or visual form of the patient's presenting complaint. For example, a problem might be introduced with the following statement:

> You are called to the emergency room of a hospital to see a 50-year old woman patient who has been rushed to the hospital after collapsing at a luncheon a half-hour ago. The patient is in severe pain.

The examinee is then required to select from a number of possibilities a course of action reflecting his estimate of the seriousness and urgency of the situation. The following are illustrative of the choices offered at this point:

You would NOW (Choose only ONE):

1. Obtain further history
2. Perform a physical examination
3. Hospitalize patient for further evaluation and therapy
4. Prepare patient for urgent surgery

   Etc.

The examinee records his decision by erasing the opaque overlay from a specially constructed answer sheet. His erasure will reveal either instructions directing him to the next section of the problem or feedback regarding the results of his decision. If, in the illustration quoted above the examinee had selected item 3, "Hospitalize patient for further evaluation and therapy", his erasure would reveal the words: "Turn to Section F." In Section F he would be confronted with an extended list of possible interventions such as those listed below, and on erasing would find the results of his orders as shown in parentheses:

In light of the available information you would NOW order (Select as many as you consider indicated):

211- Hemoglobin determination (11.0%gm%)
212- Chest X ray (see X-ray number 72)
213- Electrodardiogram (see tracing number 102)
Etc.

On the basis of these new data the examinee is required to make further decisions about the next steps in the diagnosis and treatment of this patient.

Each such problem is constructed to allow both for different medical approaches and for variation in patient responses appropriate to these several approaches. The stages in the work-up and the responses to the specific interventions the examinee chooses are meticulously designed to simulate the clinical situation. For example, in response to an order for a specific test, a laboratory report is revealed by erasure of the overlay; in response to an order for an X-ray, electroencephalogram, electrocardiogram, etc. the examinee is referred to a high quality reproduction of the X-ray or tracing; if the student orders a blood smear he is referred to a color plate of the smear; if he orders medication the patient's response is reported. Even the complications whic' must be managed differ from person to person depending (as they do in the office or clinic) on the unique configuration of prior decisions each has made. For some, the erasures will reveal an instruction to skip one or more sections of a problem because the approach they have chosen is effective in avoiding potential complications with which others must cope. If however, at any stage the examinee orders something harmful or fails to take measures essential to the recovery of the patient, he uncovers a description of the clinical features of the complication that has developed. He is then directed to a special section where he has the opportunity to take heroic measures to rectify his previous errors; if the remedial measures are inadequate he may be instructed that the problem is terminated because the patient has suffered a relapse and has been sent to another hospital or has been referred to a consultant, or has died.

The construction and analyses of these written simulations suggests that there are two somewhat distinct types: one, the diagnostic problem in which the gathering of information is the predominant element, and the other, the treatment problem in which choice among various therapeutic possibilities is the predominant element. The types of decisions required in diagnostic problems appear differ from those required in treatment problems. In the former, the choice is between doing more o. less, whereas in the treatment problems the choice is more likely to be between two mutually incompatible courses of action. The consequences of inappropriate actions will therefore also differ in the two types of problems. Most relatively short problems are easily identified as predominantly one type or the other, and the more extended problems can usually be divided into treatment and diagnostic sections and separate scores can be derived for each type of problem or from each section of the more extended problems.

Scoring Written Simulations

The problems are scored by asking a criterion group of subject matter specialists to classify each option in the problem as belonging to one of the following categories:

++ Category:	Choices which are CLEARLY INDICATED and IMPORTANT in the care of THIS patient at THIS stage in the work-up or management;
+ Category:	Choices which are CLEARLY INDICATED but of a more ROUTINE nature, i.e., should be selected but are not of special significance in the care of THIS patient at THIS stage.
0 Category:	Choices which are OPTIONAL, i.e., the probability that they will be helpful for THIS patient at THIS stage is fairly remote or quite debatable;
- Category:	Choices which are CLEARLY NOT INDICATED though NOT HARMFUL in the management of THIS patient at THIS stage;
-- Category:	Choices which are clearly CONTRA-INDICATED (i.e., are definitely harmful or carry an unjustifiablely high cost in terms of risk, pain or money) in the care of THIS patient at THIS stage.

Options in the + and ++ categories are assigned positive weights of a magnitude to reflect the importance of the decisions; Similarly, procedures the criterion group has identified as contra-indicated are given negative weights of varying sizes. The maximum number of points obtainable by selecting all indicated procedures and avoiding all useless and harmful ones is calculated. The examinee's score is reported as a percentage of this maximum. The score is called the "Net Score" or the "Proficiency Score" and is reported for a test as a whole, or for various types of problems, or for individual problems or even for sections of problems. Further it should be noted that a given Proficiency Score can be achieved in quite different ways. For example, some individuals select relatively few indicated procedures while avoiding most contra-indicated ones. Their major errors are errors of ommission, a characteristic pattern of some practicing physicians. Others select most of the indicated items, but also choose numerous contra-indicated ones. Their major errors are errors of commission, a pattern which is quite common for neophytes, such as medical students. For all special studies of written simulation the following scores have been calculated:

$$\left.\begin{array}{l}\text{Score on Proficiency}\\ \text{Score on Selection of Indicated Procedures}\\ \text{Score on Avoidance of Contra-Indicated Procedures}\end{array}\right\}\text{FOR}\left\{\begin{array}{l}\text{The Total Test}\\ \text{The Diagnostic Problems}\\ \text{The Treatment Problems}\end{array}\right.$$

Reliability of Written Simulations

The written simulation exercises are so unconventional in form that they pose new problems in defining and computing reliability.[3] In estimating the reliability of these exercises, reliability has been defined as the amount of error involved in generalizing from the results to some universe; the major issue then becomes the definition of the universe to which one wishes to generalize. All studies done to date strongly suggest that even one fairly lengthy simulation is highly reliable if one wishes to generalize to a universe of similar problems dealing with similar disease entities. However, if one wishes to generalize about some global ability, such as "clinical judgement," for example, then it is necessary to use several simulations to achieve a reasonably reliable estimate. This is explained by the fact that only modest correlations are obtained between scores on different simulation problems, for the same reasons that in "real life" lead to superb physician performance with one patient and only mediocre or even poor performance with others.

In estimating the reliability of simulation exercises for purposes of generalizing to a universe of similar problems, sampling similar components of competence, a technique analogous to the split-half method has been employed in this study. * This technique is based on the assumption that an individual has equal opportunity to select all items. This assumption is appropriate for diagnostic problems since it is possible for an examinee to make independent choices. However, in treatment problems, one choice often precludes other choices, consequently, the technique is inappropriate for most treatment problems, for which some form of analysis of variance must be used. Unfortunately, to date, no orthopedic examination has included a sufficient number of treatment problems to permit effective application of analysis of variance. For this reason, techniques of estimating reliability from corrected correlations between part test and total test scores have been utilized in this study to estimate the reliability of both diagnostic and treatment problems despite the fact that other evidence suggests that

---

* The specific method employed in this study involved obtaining the correlation between scores on every third item versus scores on the total test and correcting with Angoff Formula 12.

such a method results in a serious underestimate of the reliability of treatment problems.

The results of the various studies of reliability of written simulations are shown in Table 22 which reveals that most of the diagnostic problems achieve reliabilities in excess of .90, where one is attempting to generalize to a universe of problems similar in both content and process. Thus, we can be reasonably certain from the data that examinees who fail in a simulation exercise to diagnose a Charcot hip, for example, genuinely failed to handle that problem adequately, and that the results are not due to accidental factors. One cannot conclude, however, that such examinees would fail to make an accurate diagnosis in a clinical problem involving some other diagnostic entity; nor can one conclude that failure in one problem indicates that the physician is a poor diagnostician. Estimates of the latter must be based on analysis of performance on a large number of problems, and in a variety of settings.

Validity of Simulation Exercises

Content Validity

Though the written exercises were designed to simulate reality, they cannot duplicate it for the reasons that:

1. Confrontation with the real patient whose manner, appearance and physical reactions provide many clues, both helpful and distracting, is eliminated.

2. The exercises necessarily involve compression and distortion of time scale which may lead to exploration of blind alleys longer than is likely in the clinic or ward.

3. The examination format may impose an arbitrary pattern of exploration and intervention.

4. Real life pressures (e.g., a waiting room full of patients) are eliminated.

5. As with other examination formats, this may be preceived as requiring the examinee to anticipate what those who have constructed the examination are looking for, rather than to consider only what is best for the patient and convenient to himself.

Such challenges require thoughtful study of at least two hypotheses which underlie both written and oral simulations:

TABLE 22

RELIABILITY OF PROFICIENCY SCORES ON SIMULATION EXERCISES

EXAMINATION PROGRAM	N	PROBLEM NO.	NATURE OF PROBLEM	LENGTH OF TIME FOR PROBLEM	RELIABILITY COMPUTED BY ANGOFF 12
May 1965 Certification Exam - I	408	I	Diagnostic	45 Min.	.86
*Nov 1965 In-Training Examination	1495	I	Diagnostic	15 Min.	.53
		II	Treatment	15 Min.	.0
*Jan 1966 Certification Examination Candidates	402	I	Diagnostic	15 Min.	.57
		II	Treatment	15 Min.	-.38
		III-A	Diagnostic	15 Min.	.97
		B	Diagnostic	15 Min.	.91
		C	Treatment	15 Min.	.50
*Examiners	184	I	Diagnostic	15 Min.	.32
		II	Treatment	15 Min.	-.32
		III	Diagnostic	15 Min.	.98
		III	Diagnostic	15 Min.	.91
		III	Treatment	15 Min.	.50
May 1966 Certification Examination-I	450	I	Diagnostic	15 Min.	.91
		I	Treatment	15 Min.	-.06
		I	Total	30 Min.	.77
Nov. 1966 In-Training Examination First Year Residents	256	I	Treatment	15 Min.	.00
		I	Diagnostic	15 Min.	.97
		II	Treatment	15 Min.	.56
		II	Diagnostic	15 Min.	.81
		I+ II	Total	60 Min.	.90
Second Year Residents	464	I	Treatment	15 Min.	.23
		I	Diagnostic	15 Min.	.97
		II	Treatment	15 Min.	.40
		II	Diagnostic	15 Min.	.80
		I+ II	Total	60 Min.	.89
Third Year Residents	345	I	Treatment	15 Min.	.36
		II	Diagnostic	15 Min.	.97
		II	Treatment	15 Min.	.23
		II	Diagnostic	15 Min.	.76
		I+ II	Total	60 Min.	.89
Fourth Year Residents	390	I	Treatment	15 Min.	.56
		II	Diagnostic	15 Min.	.97
		II	Treatment	15 Min.	.39
		II	Diagnostic	15 Min.	.82
		I+ II	Total	60 Min.	.91

EXAMINATION PROGRAM	N	PROBLEM NO.	NATURE OF PROBLEM	LENGTH OF TIME FOR PROBLEM	RELIABILITY COMPUTED BY ANGOFF 12
Jan. 1968	575 **	I	Treatment	15 Min.	.00
Final		I	Diagnostic	15 Min.	.89
Certification		II	Treatment	15 Min.	-.19
Examination		II	Diagnostic	15 Min.	.76
Candidates		I+ II	Total	60 Min.	.60

* NOTE: The two problems on the November 1965 In-Training Examination were given in slightly altered form to the candidates and examiners in the 1966 Final Certification Examination.

** Sub-sample composed of all of graduates of U.S. medical schools taking the OCE for the first time.

1. That the mental processes involved in working through a simulation exercise are sufficiently similar to those required in effective clinical practice that data on the former will provide valuable information on the latter; and

2. That the individual's approach to simulation problems reveals attitudes which are characteristic of his management of actual clinical problems.

Evidence relevant to these hypotheses is presented below in the discussions of construct and concurrent validity.

Construct Validity

Several approaches to the analysis of the construct validity of written simulations have been employed in this study. The first entails investigation of the relationship between performance and such background variables as amount of training, age, and practice setting. Three hypotheses were considered in this part of the study.

(1) That increased training would be associated with higher proficiency scores and that this growth would be predominantly in therapeutic decision-making rather than in diagnostic thoroughness.

(2) That beyond a certain point, increasing age would be associated with lower proficiency scores and that this decline in performance would be manifest primarily in a greater tendency to take diagnostic shortcuts.

(3) That, since the problems had been constructed and scored primarily by physicians in academic settings, according to their value systems, clinicians in those settings would perform better than physicians practicing in other settings.

Evidence relevant to Hypothesis (1) was collected in the In-Training Examinations. Table 23 which summarizes the data on written simulations from the last three such examinations, indicates that there was no significant difference in proficiency scores on diagnostic problems between first and fourth year residents whereas, with one exception, these two groups differ significantly on proficiency scores on treatment problems. In the one exception, the treatment score was strongly linked to the diagnostic score on that particular problem. These data lead to the conclusion that increased training seems to be associated with increased ability, (as measured by these problems) to make effective treatment decisions without concomitant growth in the ability to gather information for purposes of solving diagnostic problems.

TABLE 23

Relation Between Level of Training and Proficiency Scores$^+$

Examination	Problem No.	Type of Problem	Mean Proficiency Score of Residents In: First Year	Second Year	Third Year	Fourth Year	Difference Between First and Fourth Year
Nov. 1965 In-Training Examination	I	Diagnostic	73	71	74	75	+ 2
	II	Treatment	28	31	31	35	+ 7*
			258	309	369	430	
Nov. 1966 In-Training Examination	I	Diagnostic	62	63	55	59	- 3
	II	Treatment	-10	- 7	- 7	- 4	+ 6*
	II	Diagnostic	7	7	6	5	- 2
	II	Treatment	17	20	20	15	- 2
			456	531	345	390	
Nov. 1967 In-Training Examination	I+II+III	Diagnostic	36	36	36	36	0
	I+II+III	Treatment	43	46	52	54	+11*
			244	513	499	399	

+ All scores are expressed as raw scores.

* Significant at .05 level of confidence.

Insight into the possible explanations of these phenomena was obtained from a study of the responses of residents, candidates and examiners to similar problems. These data, summarized in Table 24, suggest that those with more experience tend to be more willing to trust their judgement and to take unavoidable, radical action earlier than less experienced physicians.

TABLE 24

COMPARISON OF MANAGEMENT DECISIONS AMONG SELECTED GROUPS

1966 In-Training and Certification Examinations

Recommended Action	Percentage of Each Group Selecting the Option		
	Residents (N=1366)	Candidates (N=403)	Examiners (N=184)
Amputate, * first opportunity	10	20	28
Amputate, * later	26	20	22
Total	36	40	50

* Amputation was the optimal course of action.

Data on hypotheses (2) and (3) were obtained from the 1966 Final Certification Examination in which identical written simulation exercises were administered to both candidates and examiners, and the responses of the latter group were further analyzed according to age and academic affiliation. The results, summarized in Table 25, reveal that, among the examiner group, increasing age was associated with lower scores on both treatment and diagnostic problems and that, while the younger examiners showed a marked superiority to the candidates on the treatment problems, no such superiority was manifest on the diagnostic problems. Table 26 suggests one possible explanation: In the diagnostic work-up, examiners seek substantially less information than candidates, and as indicated in Table 27, this tendency among examiners is exacerbated with increasing age, a finding similar to those noted in the observational studies of Peterson and Clute.[4,5] Finally, as shown in Table 25, the responses of full-time academicians on the written simulations were in much closer agreement with the criterion group than were the responses of examiners from other practice settings. In short, the experienced physician is more likely to take diagnostic shortcuts and is more willing to take decisive action in treatment; among the experienced physicians, those who practice in academic settings are, not surprisingly, more likely to behave according to standards established by criterion groups which are heavily weighted with academicians.

In summary, the data from studies of the construct validity of the written simulations are encouraging in that differences in the responses of various groups are in the expected direction and the patterns of response are closely similar to those reported in observational studies of physician performance in clinical settings.

TABLE 25

PROFICIENCY SCORES OF EXAMINERS AND CANDIDATES ON WRITTEN SIMULATIONS
1966 FINAL CERTIFICATION EXAMINATION

	Candidates	Examiners						
		By Age Groups:			By Affiliation:			Total
		37-45	46-55	56-65	Full-time Academic	Part-time Academic	Other	
	N=403	N=55	N=63	N=42	N=40	78	42	160
Problem I Diagnostic (Laboratory)	71	73	70	62	81	64	66	69
Problem II Treatment	22	34	21	22	36	24	19	
Problem III Diagnostic (History and Physical Sections)	29	29	21	10	28	17	22	21
III Diagnostic (Laboratory Section)	41	34	37	22	38	29	32	32
III Treatment	23	32	18	6	33	17	12	19
Total +	35	38**	29**	21**	40*	27*	26*	30

+ The total score was obtained by weighting Problem I as 1/9, Problem II as 2/9 and each section of Problem III as 2/9.

* Differences significant .05 level

** Differences significant at .01 level

TABLE 26

ANALYSIS OF STRATEGIES IN GATHERING INFORMATION ON WRITTEN SIMULATIONS

1966 Final Certification Examination, Problem III

Procedure	Percentage Selecting as the Initial Procedure		Total Percentage Selecti Procedure at Any Point	
	Candidates	Examiners	Candidates	Examiners
* Obtain a history	61	42	67	50
Obtain a physical examination	8	13	66	57
Obtain laboratory and X-ray data	10	12	64	48
Obtain a biopsy	19	29	95	96
Initiate treatment	2	4	92	92

* Optimum choice according to criterion group

TABLE 27

EXAMINER'S PROFICIENCY SCORES ON HISTORY AND PHYSICAL EXAMINATION SECTION OF WRITTEN SIMULATIONS

1966 Final Certification Examination, Problem III

Age Group	N	Mean %	Standard Deviation
Under 40	16	38.3	31.1
41-45	44	25.3	25.2
46-50	36	22.1	24.4
51-55	32	19.8	23.2
56-60	33	15.1	20.7
over 60	18	1.3	3.8

Evidence regarding the construct validity of written simulation exercise was also obtained from correlational and factor analytic studies of the relationships between performance on the simulation exercises and that on other types of test exercises. Since the written simulation exercises had been designed to measure aspects of competence not sampled by more conventional devices it was predicted: (1) that there would be relatively low correlations between scores on written simulation exercises and scores on other tests and (2) that the factor structure of scores on written simulations would differ from that of scores on other tests.

Data relevant to the first hypothesis are presented in Table 28. These data, as accumulated from the administration of a number of problems in several different tests given to different groups of examinees, are consistent: The correlation between scores on written simulation exercises and scores on other evaluative techniques is in no case high and in most cases does not differ from zero. The data suggest that not more than 10%-20% of the variance is common to all techniques and that this common variance can probably be attributed to a common informational base requisite to performance on any of the tests. Beyond that, the orals, the simulation exercises and the multiple choice questions appear to be measuring somewhat different aspects of competence.

Data relevant to the second hypothesis are presented in Tables 6 and 7 above which summarize the more significant results obtained in the two factor analytic studies that have been conducted to date. In the first study as summarized in Table 6 three independent factors emerged: The multiple choice examination and conventional orals loaded on one factor; written simulations of the diagnostic type loaded heavily on a second factor; and written simulations of the treatment type loaded moderately on a third factor. The second study, summarized in Table 7, revealed a somewhat more complex, but essentially similar factor structure in the 1968 Final Certifying Examination. In short, both yield data compatable with the second hypothesis stated above.

Concurrent-Validity

The concurrent validity of the written simulation exercises was investigated by means of correlational and multiple regression analyses in which total and sub-scores on the exercises were used together with other test variables to predict supervisor's ratings of the habitual performance of residents. In the first such study, conducted on the 1966 In-Training Examination, residents in the first 2 years of the training program and those in the last 2 years were treated as different populations since it was felt that supervisors would use different criteria in evaluating the two groups. The results, as summarized in Table 29 and detailed in Appendix 14, indicate that the total score on the Written Simulation Exercises makes no significant contributions to the prediction of supervisors' ratings of any aspect of habitual performance. However, the picture differs markedly when sub-score on the

TABLE 28

CORRELATIONS BETWEEN WRITTEN SIMULATION PROFICIENCY SCORES AND SCORES ON OTHER EVALUATION TECHNUQIES

PROBLEM NO.	TYPES OF SCORES	ADULT CONVENTIONAL ORAL (½ Hr)	PROBLEM SOLVING ORAL (15 MIN)	SIMULATION ORAL (10 MIN)	MULTIPLE CHOICE
January 1966 Final Cert Exam N=383					
I	Diagnosis (Lab)	.14	.05	.05	.24
II	Treatment	-.03	-.07	-.01	.07
III	Diagnosis (Hist. & Phys.)	.05	.12	.06	.10
	Diagnosis (Lab)	.18	.10	.11	.18
	Treatment	.03	.00	.06	.08
May 1966 Cert. Exam I N=408					
I	Diagnosis	-	-	-	.16
II	Treatment	-	-	-	.18
I	Total	-	-	-	.21
November 1966 In Training Exam First Two Years N=109					
I	Treatment	.08	-.02	-.01	.01
I	Diagnosis	-.05	+.24	+.12	.07
II	Treatment	-.08	-.05	-.07	.07
II	Diagnosis	-.19	+.10	+.17	.06
I&II	Total	-.06	.11	-.01	.05
November 1966 In-Training Exam Second Two Years N=109					
I	Treatment	.19	.07	-.00	.23
I	Diagnosis	-.14	.04	-.00	.23
II	Treatment	+.14	-.06	-.01	-.05
II	Diagnosis	-.17	.07	-.15	-.04
	Total	-.06	.05	-.04	.18
January 1967 Cert Exam N=407					
I	Treatment	.01	.02	.05	.08
November 1967 In-Training Exam N=1682					
I+II+III	Treatment	-	-	-	.27
I+II+III	Diagnosis	-	-	-	.29
January 1968 Final Cert Examination N=784			*		
I+II	Diagnosis	-	.16	.02	.31
I+II	Treatment	-	.20	.07	.24
I+II	Total	-	.23	.04	.35

*NOTE: In the January 1968 Examination the Problem Solving Scores are based on a combination of 4 half-hour orals.

simulation exercises are considered. For example, for residents in the first two years of training, scores on the avoidance of contra-indicated procedures in diagnostic problems are negatively correlated with supervisors' ratings. That is, individuals who ask numerous irrelevant questions and who select many contra-indicated procedures seem to be more inquisitive and more thorough though probably less well-informed than others in their diagnostic inquiries, and it appears that the chief of training values curiosity and thoroughness and tends to disregard lack of information on the part of the relatively inexperienced residents. In contrast, for residents in the last two years of training scores on the avoidance of contra-indicated procedures in diagnostic problems were positively correlated with supervisors' ratings, probably a reflection of the differences in standards training chiefs apply to neophytes and to experienced practitioners.

Such data raised so many puzzling questions that the concurrent validation study of the 1966 In-Training Examination was replicated on the 1968 Final Certifying Examination. The results are summarized in Table 30 and detailed in Appendix 15 : They suggest, as did the study of the 1966 In-Training Examination, that several different types of tests are needed to predict competence as defined by supervisors, since in all cases the three best predictor included scores from both written and oral exercises. Second, as would be expected, the several predictor variables are differentally useful in predicting different criterion behaviors: For example, the partial correlation of the Multiple Choice Recall score with supervisors' ratings of Information Gathering Behavior is substantially higher than that with ratings of Effectiveness in Emergency Care. Third, certain of the subscores in the simulation exercises appear to make a useful contribution to the prediction of supervisors' ratings of certain types of affective behavior, for example, Effectiveness in Colleague and in Patient Relationships, in assuming Continuing Responsibility and in providing Emergency Care. In short, these data suggest that the written simulations may be measuring certain styles and attitude sets as well as sampling purely cognitive behavior.

Summary

In summary, studies of the reliability and validity of the written simulation exercises suggest that such problems do require the examinees to demonstrate certain types of behavior similar to that required of them in clinical settings. However, the ability to handle such problems differs markedly from one problem to another depending in part on the content of the problem, hence, several problems are required to obtain reasonably reliable results. Nevertheless, despite limited generalizability across problems, scores based on even a few problems, when combined with data from other evaluation techniques, make a significant contribution to the assessment of competence in orthopaedics. Secondly, such exercises appear to sample some types of behavior (e.g. diagnostic thoroughness) not easily observed by superviosrs and not generally rewarded in training programs. Third, exercises of this type appear to sample certain affec ive components of completence not readily measured with conventional cognitive techniques.

Though further work is required to develop optimum methods of constructing and scoring simulation exercises they appear to be making a sufficiently reliable and valid contribution to the more comprehensive assessment of professional competence to justify their expanded use.

1. R. Damrin, Glaser, *etal*, "The Tab Item: A Technique for the Measurement of Proficiency in the Problem Solving Task", A. A. Lumsdaine and R. Glaser (Eds), *Teaching Machines and Programmed Learning: A Source Book*, Washington, NEA, 1960, pp. 275-285.

2. Christine McGuire and David Babbott,"Simulation Technique in the Measurement of Problem Solving Skills", *Journal of Educational Measurement*, Spring, 1967, pp. 2-12.

3. Arieh Lewy and Christine McGuire, "A Study of Alternate Approaches in Estimating the Reliability of Unconventional Tests", Read at Annual meeting of the AERA, Feb. 18, 1966.

4. Kenneth Clute, *The General Practionier*, (Toronto) University of Torondo Press, 1963.

5. Osler Peterson, *etal*, "An Analytical Study of North Carolina General Practice", 1953-54, *Journal of Medical Education*, 1956, 31, No. 12 (whole part 2).

TABLE 29-A

RESULTS OF MULTIPLE CORRELATIONAL ANALYSIS USING RATING FACTORS AS DEPENDENT VARIABLES AND TEST SUBSCORES AS INDEPENDENT VARIABLES
1966 In-Training Examination

FIRST AND SECOND YEAR RESIDENTS
N = 109

Dependent Variables: Rating Factors	R	F	Independent Variables Test Scores	Partial r	F
Factual Information	.53	1.83**	Problem I, Written Simulation Score on Avoidance of Contra-Indicated Diagnostic Procedures	.28	7.84**
			Proposed Treatment Interview Overall Score	.28	7.42**
			Proposed Treatment Interview Interaction Score	-.26	6.46*
			Problem II. Written Simulation Score on Avoidance of Contra-Indicated Treatment Procedures	.24	5.20*
			Problem I, Written Simulation Score on Selection of Indicated Diagnostic Procedures	-.22	4.71*
Problem Solving			None Significant		
Information Gathering			None Significant		
thics			None Significant		

* Significant at .05 level
** Significant at .01 level

TABLE 29-B

RESULTS OF MULTIPLE CORRELATIONAL ANALYSIS USING RATING FACTORS AS DEPENDENT VARIABLES AND TEST SUBSCORES AS INDEPENDENT VARIABLES
1966 In-Training Examination

THIRD AND FOURTH YEAR RESIDENTS
N = 119

Dependent Variables: Rating Factors	R	F	Independent Variables Test Scores	Partial r	F
Factual Information			None Significant		
Problem Solving	.56	1.85*	Adult Oral Overall	.22	4.87*
			Problem I, Written Simulation Score on Avoidance of Contra-Indicated Diagnostic Procedures	.20	4.02*
Information Gathering	.63	2.70*	Diagnostic Interview Subscore on Diagnosis	.21	4.41*
			Problem I, Written Simulation Score on Avoidance of Contra-Indicated Treatment Procedures	.20	3.85
Ethics	.55	2.01*	Proposed Treatment Interview Subscore on Manner	-.31	9.14**
			Problem II, Written Simulation Score on Selection of Indicated Treatment Procedures	.28	7.28**
			Problem I, Written Simulation Score on Selection of Indicated Diagnostic Procedures	-.21	4.0[illegible]

* Significant at .05 level
** Significant at .01 level

TABLE 30

RESULTS OF MULTIPLE REGRESSION ANALYSIS USING RATING FACTORS AS DEPENDENT VARIABLES AND TEST SCORES AS INDEPENDENT VARIABLES

1968 FINAL CERTIFYING EXAMINATION

Dependent Variables: Rating Factors	Reliability of Dependent Variable	R	F	Independent Variables: Test Scores	Partial r	F
Information Gathering	.29	.36	5.13**	Multiple Choice-Recall	.12	4.48*
				Observation and Interpretation Oral-Interpretation	.11	4.48*
				Trauma Oral-Problem Solving	.10	3.84
Problem Solving	.29	.40	6.72**	Observation and Interpretation Oral-Interpretation	.16	9.81**
				Multiple Choice-Recall	.14	7.56**
				Simulation Oral-Attitudes	.09	3.31
Clinical Judgment	.22	.34	4.46**	Observation and Interpretation Oral-Interpretation	.13	6.21*
Surgical Technique	.29	.26	2.42**	Observation and Interpretation Oral-Interpretation	.12	5.39*
Patient Relations	.25	.27	2.82**	Simulation Oral-Attitudes	.11	4.35*
				Written Simulation-Diagnostic	.11	4.20*
				Observation and Interpretation Oral-Interpretation	.10	3.78*

TABLE 30 (con't)

Dependent Variables: Rating Factors	Reliability of Dependent Variable	R	F	Independent Variables: Test Scores	Partial	F
Continuing Responsibility	.23	.29	3.30**	Observation and Interpretation Oral-Interpretation	.10	3.54
				Written Simulation-Diagnostic	.09	3.42
Emergency Care	.28	.28	3.24**	Observation and Interpretation Oral-Interpretation	.15	8.12**
Colleague Relation	.28	.26	2.50**	Observation and Interpretation Oral-Interpretation	.10	3.31
Ethics	.28	.25	2.41**	Adult Oral-Problem Solving	.09	3.05
Overall Competence	.31	.37	2.80**	Observation and Interpretation Oral-Interpretation	.12	5.48*
				Multiple Choice-Recall	.12	5.46*
				Multiple Choice-Problem Solving	.09	3.13

*Significant at .05 level
**Significant at .01 level

CHAPTER VII

ORAL EXAMINATIONS

Conventional oral examinations are increasingly subject to the often legitimate criticisms that they are inherently unreliable due to both sampling and rating errors, that they are frequently invalid because they are not designed to evaluate a number of important areas of competence, that the aspects of competence which they do evaluate cannot be precisely determined because they are so unstructured and unstandardized and that they are unduly expensive in terms of examiner time and administrative effort. Despite these defects the oral examination has persisted as an evaluation technique in those situations where the limited number of examinees permit, primarily because the stubborn conviction prevails that the oral examination measures some ill-defined aspects of performance not measured by other means, and because the oral preserves some element of personal contact as a basis for making an important decision about an individual and also provides the examiner with a feeling of participation in the evaluation process difficult to obtain with other methods. Consequently, in restructuring the traditional oral to avoid the defects and capitalize on the benefits cited above, two considerations were paramount in the present study. The first was to utilize the opportunity afforded by an oral examination to sample such interpersonal skills as ability to relate to and communicate with patients and colleagues. The second was to take advantage of opportunities for examiner-examinee dialogue to sample the higher level cognitive processes entailed in interpretation and problem-solving. These considerations led to the development of three quite different examination formats, each deserving of separate analysis.

The Overall Design of the New Oral Examinations

Previous experience with written simulations had confirmed numerous advantages in sampling complex cognitive processes with exercises that require the examinee to make decisions similar to those required in real life; however, such exercises in written form have certain limitations. Specifically, they permit the examinee to select from a list of possible responses rather than requiring him to generate his own inquiries; in so doing they impose certain strictures on the examinee's mode of inquiry that are not present in real life, while relieving him him of certain pressures toward efficiency that are characteristic of the clinical situation; finally, in a written exercise, it is always possible to know what a person chooses but often impossible to determine why he makes a particular choice.

It appeared to the study staff that problem-solving exercises could be developed in an oral format that retained the essential characteristics of the written simulation exercises while avoiding their limitations.

In the first such exercise, the Diagnostic Interview, the examinee was given a brief description of a patient's presenting complaint. It was his task to question the examiner, who was programmed with the details of the case, in order to arrive at a diagnosis. During the history-taking part of the inquiry the examinee played the role of physician while the examiner played the role of a patient; no role-playing was involved in the inquiry regarding physical and laboratory findings. The examinee was allowed 12 minutes to gather the information and three minutes to explain his diagnostic impressions. A second examiner observed the proceedings but did not participate.

This exercise, together with two other role-playing exercises designed to evaluate ability to relate to and communicate with patients and with colleagues, was first administered experimentally in January, 1966 along with four conventional oral examinations. In order to maximize the validity and reliability of these experimental orals, standardized case materials were prepared, an objectified rating form was developed (see Appendix 10) and all participating examiners were asked to attend a two day training program to prepare them for administering and scoring the new examinations.

Analysis of the results of these experimental examinations together with those obtained from a second experimental series conducted in 1967 led to the following restructuring of the oral component of the 1968 Final Certification Examination: The 2½ hours alloted to the orals was divided into 5 half-hour examinations one of which was designed to sample skills in relating effectively to patients and colleagues, one to sample interpretive skills and the remaining 3 to sample problem-solving skills. Of the three problem-solving orals, one was devoted to problems of adult orthopedics, one to problems of children's orthopedics and one to problems of trauma.

The half-hour designed to sample skills in relating to patients and colleagues consisted of 3-4 role-playing exercises in which the candidate took the role of physician and the examiner took the role of patient or colleague in a specifically described situation typical of the problems the physician encounters in working with others.

The nature of the half-hour oral designed to sample interpretive skills is probably best conveyed by the instructions to examiners shown in the exhibit on the following page. During this exercise the candidate was presented with 3-4 sets of slides and/or X-rays, each set relating to a single case. It was his task to describe the findings present.

Each of the half-hour orals devoted to problem-solving consisted of 2-3 problems of the following types:

1. The Diagnostic Problem: This type of problem resembled the Diagnostic Interview described previously with the single exception that role playing was entirely eliminated as included in the history-taking portion of the earlier format.

2. The Defense of Therapy Problem: In this type of problem the examinee was presented with a brief description of a specific clinical case; it was his task to outline in detail the plan of management he would recommend; it was the examiner's task to ask probing questions that would require the candidate to explain his rationale and to defend his decisions.

3. The Emergency Treatment Problem: In this type of problem the examinee was presented with a brief description of a specific emergency case; it was his task to detail the steps he would take and to indicate his priorities in that particular emergency, while the examiner provided feedback regarding the consequences of each action recommended by the examinee.

4. The Complication Problem: In this type of problem the examinee was presented with a brief description of a specific case representing a problem of chronic illness. As in the Emergency Treatment Problem it was his task to detail the steps he would take while the examiner provided feedback regarding the consequences of each action he recommended.

The first two types of problems were employed in the problem-solving orals on Adult and Children Orthopedics, and the latter two types in that on Trauma.

In all 5 oral examinations the candidate was rated on 4 separate factors, i.e. components of competence; these were recall of factual information, analysis and interpretation of clinical data, problem-solving ability and professional attitudes as defined in Table 31. While the same factors were considered in scoring all examinations, the value assigned each factor varied among the 5 orals so as to place maximum weight on the one or more aspects of competence each type of oral was specifically designed to sample.

EXHIBIT IV

Administration of the Observation and Interpretation Examination

You should first present the material to the candidate and instruct him to describe what he sees precisely as he might in a written report, indicating any abnormalities that may be present.

If the candidate fails to interpret the material properly you might supply him with some additional historical, physical examination or laboratory data which would assist him. You should not provide this additional information until AFTER he has initially described his findings. If he identifies some abnormalities you should then ask additional questions which would probe his ability to interpret what he sees in the light of his knowledge and understanding of physiological and pathological processes. For example, you might ask him to speculate as to the reason that the structures on the slide show the pattern they do, or you might ask the probable effect on the abnormality of various types of treatment.

DO NOT SPEND TOO MUCH TIME ON ANY ONE EXERCISE.
Remember, you are mainly concerned with what the candidate sees and how he interprets it. If you ask too many questions about diagnosis the candidate may simply answer them on the basis of his basic information and not on the basis of any observational skills. Although we recognize that it is extremely important to assess the candidate's basic store of information on diagnosis and treatment, this area of competence is being probed thoroughly in other portions of the examination.

TABLE 31

Oral Examination Rating Form-Explanation of Factors

Factor I Recall of Factual Information

This factor deals with the candidate's factual knowledge of general medicine and orthopaedics as displayed by his ability to discuss the cases during the examination. Note that candidates can score high on this factor, but still do badly in Factor III because they have difficulty in integrating the information they possess. If you believe that the factual content of the examination is too simple to allow for any judgment of the store of information of the candidate, check "Unable to Judge" rather than "Good" or "Excellent".

Factor II Analysis and Interpretation of Clinical Data

This factor deals with the candidate's ability to perceive the characteristics, both normal and abnormal, of material presented to him in visual form (such as X-rays, slides, motion pictures, photographs, etc.) and to explain what he has seen.

Factor III Problem-Solving Ability--Clinical Judgment

This factor deals with the candidate's ability to use the information he has to make appropriate decisions in patient diagnosis and treatment as displayed by the data he solicits about patients, the diagnostic and therapeutic conclusions he comes to, his ability to provide a rationale for the decisions he makes.

Factor IV Relates Effectively--Shows Desirable Attitudes

This factor deals with the ability of the candidate both in statements and in manner to communicate effectively and convey genuine concern for patients, respect for colleagues and an understanding of the ethical responsibilities of a physician in his relationships with others.

In order to ensure that all examinations were administered and scored as planned, detailed instructions to examiners and to candidates were preposed outlining the procedures to be followed (see Appendices 16 and 17): Secondly, as in the previous experimental orals, training sessions were conducted for all examiners and standardized case materials were supplied to each. Finally, to ensure maximum objectivity and reliability of the orals, examiners were instructed to utilize a 12-point scale in rating candidate performance*, and to record their judgments on the special form reproduced below.

ORAL EXAMINATION RATING FORM

	Unable To Evaluate	Definite Failure	Marginal	Good	Excellent
Recall of Factual Information	☐ 00	☐ ☐ ☐ 01 02 03	☐ ☐ ☐ 04 05 06	☐ ☐ ☐ 07 08 09	☐ ☐ ☐ 10 11 12
Analysis and Interpretation of Clinical Data	☐ 00	☐ ☐ ☐ 01 02 03	☐ ☐ ☐ 04 05 06	☐ ☐ ☐ 07 08 09	☐ ☐ ☐ 10 11 12
Problem-Solving Ability; Clinical Judgment	☐ 00	☐ ☐ ☐ 01 02 03	☐ ☐ ☐ 04 05 06	☐ ☐ ☐ 07 08 09	☐ ☐ ☐ 10 11 12
Relates Effectively; Shows Desirable Attitudes	☐ 00	☐ ☐ ☐ 01 02 03	☐ ☐ ☐ 04 05 06	☐ ☐ ☐ 07 08 09	☐ ☐ ☐ 10 11 12

Data on the reliability and validity of each of the new oral examinations are summarized in the succeeding sections of this chapter.

---

*Analogous factor scores derived from the multiple choice and written simulations were converted to the same 12-point scale to facilitate comparison of factor scores across techniques and computation of a composite score on each factor.

Reliability of the Oral Examinations of Complex Cognitive Skills

The Diagnostic Interview

The interrater reliability of the Diagnostic Interview was assessed by having two examiners independently rate an examinee on the same fifteen minute interview. The results of these studies, shown in Table 32A, indicate high inter-rater agreement, especially in the light of the limited size of the observational sample. However, prior experience with written simulations suggested that adequate reliability across cases would be much harder to achieve, and in the sole study of the combined effects of error, due both to examiner disagreement and to variation in examinee ability to handle different cases, the estimated coefficient of reliability was very low -- .25 (see Table 32B). Since most of the error appeared to be attributable to restrictions in the content sampled, it was concluded, that, if the oral examinations were to be useful evaluation techniques, it would be necessary to increase the number of cases to which each candidate was exposed and to pool the data from all examiners and all problems. This procedure was followed in the 1968 Final Certification Examination with the results reported below.

The Oral Tests of Complex Cognitive Skills as Revised

Due to administrative complications, it was impossible to obtain estimates of interrater reliability of the oral problem-solving and interpretation exercises because only one examiner was available to administer each examination. However, if one assumes that the four cognitive orals employed in the 1968 Final Certification Examination are equivalent forms, then the correlations between them can be used as estimates of reliability. Table 33A reports the intercorrelation of the Problem-Solving scores on the four examinations. The Spearman-Brown correction formula yields a reliability estimate of .47 for a combination of four tests with an average reliability of .18. These results are consistent with those obtained by employing the ANOVA formula developed by Ebel[1], in which each group of candidates who were rated by the same team of examiners is considered as a block. As shown in Table 33B, the reliability estimates for the 4 such blocks studied in the 1968 Final Certification Examination ranged from .40 - .63 and the average was .53. The results from the two methods of estimating reliability indicate that the best estimate of the combined sampling and rater reliability of the oral problem-solving score on that examination was approximately .50.

TABLE 32A

INTERRATER RELIABILITY IN SCORING OVERALL COMPETENCE ON DIAGNOSTIC INTERVIEWS

Examination	N	Mean Scores	Correlation of Team Scores	Reliability *
1966 Final Cert. Exam.	383	6.7	.58	.73
1967 Final Cert. Exam.	387	6.8	.63	.78
1966 In-Training	33	Not Computed	.63	.78

* Computed by Spearman-Brown formula for pooled scores of both examiners.

TABLE 32B

RELIABILITY OF THE OVERALL COMPETENCE SCORE ON THE DIAGNOSTIC INTERVIEW ACROSS CASES AND EXAMINERS

Examination	N	Correlations with Another Examiner Using a New Case	Reliability *
1966 In-Training	25	.14	.25

* Computed by Spearman-Brown formula for pooled scores of both examiners.

TABLE 33A

INTERCORRELATIONS OF PROBLEM SOLVING SCORES ON THE THREE PROBLEM SOLVING ORALS AND THE OBSERVATION AND INTERPRETATION ORAL, 1968 FINAL CERTIFICATION EXAMINATION

N=391

	Adult	Child	Trauma	Observation and Interpretation
Adult	---	.17	.29	.13
Child	.17	---	.18	.18
Trauma	.29	.18	---	.12
Observation and Interpretation	.13	.18	.12	---

TABLE 33B

ESTIMATES OF RELIABILITY OF ORAL PROBLEM SOLVING SCORES
1968 FINAL CERTIFYING EXAMINATION

Block	N	Reliability of One Score in Block	Reliability of Pooled Scores in Block
A	27	.29	.63
B	27	.14	.40
C	27	.23	.55
D	27	.22	.53
Mean		.22	.53

While pooled data from 4 oral examinations may be extremely useful in generalizing about groups, they are not sufficiently reliable to use alone to certify individuals; to be of value for this purpose, they must be used in combination with other test data. As shown in the next section, the data on the validity of the oral examinations support this view.

Validity of the Oral Examinations of Complex Cognitive Skills

Content Validity

The process analysis of the conventional orals revealed that they measured predominantly the recall of factual information. The new orals were deliberately designed to elicit more complex cognitive behavior from candidates. However, there was considerable question as to whether the examiners, accustomed as they were to administering oral quizzes, would be able to adjust to the new examination techniques. Consequently, a systematic observational analysis of a random sample of the over 3500 oral examinations administered in January 1968 was made by a trained team composed of 12 physicians and educators.* The results (see Table 34) indicate a significant shift in the behavior of both examiners and candidates: in the traditional orals, candidates spent most of the time replying to specific questions posed by the examiner; in contrast, in the new orals they spent most of the time questioning the examiner to obtain data for interpretation in arriving at conclusions which they then explained to the examiner. Secondly,

---

* See Appendix 28 for the complete report of the observational analysis.

TABLE 34: NATURE OF EXAMINER-CANDIDATE BEHAVIOR

Type of Examination	Number of Observations	Percent of Total Behavioral Units Recorded As:																		Average No. of Interactions per half-hour	Average No. of discrete topics or problem-situations per half-hour
		Examiner Behavior Category:*									Candidate Behavior Category:**										
		1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9		
I-Simulation	53 (8 duplicate)	7	9	1	2	6	1	1	12	10	1	6	15	3	10	1	0	2	15	46	2.7
II-Observation and Interpretation	45 (9 duplicate)	6	27	3	3	5	2	4	1	0	1	3	8	19	15	2	1	0	0	49	2.7
III-Problem-Solving	137 (24 duplicate)	4	16	2	2	18	2	2	1	1	1	19	6	3	20	2	1	0	1	58	1.8

* Types of Examiner Behavior

1	Asks for information
2	Asks for specific interpretation, conclusion or recommended action
3	Asks for reasons, evidence or criteria
4	Gives general instruction or information
5	Gives specific data
6	Gives clarification or interpretation
7	Provides cues
8	Expresses hostility or challenge
9	Asks for reassurance or understanding

** Types of Candidate Behavior

1	Requests general information or clarification
2	Requests specific data
3	Gives information, generalization (recall)
4	Gives specific observation
5	Gives intervention, inference, summary conclusions
6	Gives principle or reason on demand
7	Gives evidence of impirical validation on demand
8	Expresses hostility, rejection
9	Persuades, influences, manipulates, reassures

NOTE: The meanings of Category 1 of examiner behavior and Category 3 of candidate behavior differ in the simulation (I) and non-simulation (II and III) examinations. In the former, an examiner, playing the role of a patient may request information about "his" illness (to be coded 1 or 9 depending on the predominant element in the request), and the candidate may respond with specific information (coded 3) or with reassurance (coded 9) or may ignore the plea for help (coded 8). In the non-simulation examinations these symbols take on a conventional meaning applicable to any type of oral examination.

examiner-candidate behavior differed in the expected directions among the three major types of new orals (problem-solving, interpretation and role-playing).

As a further check on the content validity of the examinations, examiners and candidates were asked to complete a questionnaire indicating their acceptance of the new techniques.* Table 35 reports examiners' responses to the statement: "The portion of the examination I administered provided me with valuable information about the candidate's ability in some important area of orthopaedics."

TABLE 35

EXAMINER APPROVAL OF NEW ORAL EXAMINATIONS
1968 FINAL CERTIFICATION

Examination	N	Strongly Agree		Agree		Undecided		Disagree		Strongly Disagree	
		N	%	N	%	N	%	N	%	N	%
Problem-Solving: Adult	36	6	17	30	83	0	-	0	-	0	-
Problem-Solving: Child	35	4	11	29	83	1	3	1	3	0	
Problem-Solving: Trauma	34	9	26	25	73	0	-	0	-	0	-
Observation and Interpretation	34	14	41	18	53	2	6	0		0	
Role-Playing Simulations	50	11	22	28	56	4	8	5	10	2	4
Total	189	94	23	130	68	7	4	6	3	2	1

While the results reveal some reservations about the value of the role-playing simulations, they also indicate an overwhelming examiner approval of the new cognitive orals:

---

* See Appendices 18 and 19 for the complete report of the questionnaire study.

Among the over 100 examiners who administered the new problem-solving orals, only one felt that they did not yield valuable information about an important aspect of competence and only one other was undecided about this issue. Candidate response, as shown in Table 36, was almost equally favorable. Of special interest in this regard, is the fact that there were no significant differences between the responses of candidates who passed the examination and those who failed it.

TABLE 36

PERCENT OF CANDIDATES AGREEING WITH SELECTED STATEMENTS ABOUT COMPONENTS OF THE 1968 CERTIFYING EXAMINATION

Statement	Response Group	Written Tests		Oral Examinations				
		Multiple Choice Test	Written Simulations	Problem-Solving Orals			Observation and Interpretation	Role-Playing Simulations
				Trauma	Adult	Childrens		
Gave me a chance to demonstrate my abilities in some important areas of orthpaedic surgery	100 who passed	43	61	86	75	82	69	67
	100 who failed	42	52	85	70	75	65	65
Most topics covered were important in orthopaedic practice	100 who passed	52	76	91	85	86	72	78
	100 who failed	52	79	91	84	87	73	77
Examiner was skillful in putting me at my ease	100 who passed			72	68	74	72	78
	100 who failed			63	60	66	64	70
Examiner DID NOT give me a chance to answer questions adequately	100 who passed			4	11	3	2	
	100 who failed			3	3	1	1	0

In summary, the data from both systematic observational analysis and from questionnaire studies suggest that the new orals are characterized by high content validity with respect both to the material sampled and the examinee behavior elicited.

Construct Validity

To determine the construct validity of the new examinations, data from the perceptions of observers, candidates and examiners were supplemented, where possible, by studies of examinee performance at different levels of education and experience, and by factor analytic studies of the interrelations among various test scores. Data of the first type are available only for the Diagnostic Interview, and as summarized in Table 37, indicate that differences between groups at different levels of training are statistically significant and in the expected direction.

TABLE 37

OVERALL COMPETENCE SCORE ON DIAGNOSTIC
INTERVIEW BY LEVEL OF TRAINING

Level of Training	N	Scores Mean*	S.D.
1st year	29	5.4	2.2
2nd year	75	6.8	1.7
3rd year	50	6.9	2.5
4th year	79	7.6	2.5
Total	233	6.9	2.8

* Differences between means significant at .01 level by ANOVA

Data from the two factor analytic studies bearing on the construct validity of the new oral examinations are summarized below.*

---

* See Tables 6 and 7 and Appendices 14 and 15 for detailed presentation of data.

In the 1966 Final Certification Examination, the Diagnostic Interview was shown to have a moderate loading on a recall or content factor, and a heavy loading on a separate factor. This second factor appeared to be related to the ability to withhold judgment, since it was at the opposite pole from the Proficiency Score on the written simulations of treatment problems included in that examination which placed emphasis on decisiveness in taking action. The factor analytic study of the 1968 Final Certification Examination indicated that restructuring the oral examinations altered the factor structure in the direction of significantly increased factorial complexity. In commenting on these data, it should be noted, first that all of the orals had at least moderate loadings on a factor which appears to be the ability to respond effectively in oral situations; this factor was so-named because the Role-playing Simulations, which are designed to test ability to communicate with and relate to others, had the highest loading on this factor. Second, each of the four new cognitive orals had at least moderate loadings on two or more factors, and one, (the Observation and Interpretation Oral) had loadings on all five of the factors that emerged from the analysis. Two of the Problem-Solving Orals, Trauma and Adult Orthopaedics, had substantial loadings on a content factor; the third, Children's, did not load at all on this factor. The Observation and Interpretation Oral loaded moderately on this factor, and both it and the Problem-Solving Oral in Children's Orthopaedics had moderate loadings on a factor which appeared to be related to the ability to draw reasonable inferences from ambiguous data; this factor was so named because the score on Selecting Indicated Procedures in the Written Simulations of treatment problems also had a very high loading on it.

In summary, the factor analytic studies strongly suggested that at least some of the new orals were multi-factorial in nature. In recognition of this fact, all oral examiners were instructed to rate candidates on several aspects of competence and these sub-scores were differentially weighted in the various oral examinations, so as to assign greatest importance to the one component which each new type of exercise had been specifically designed to assess. While it was recognized that the intercorrelations among these ratings could be expected to be relatively high, it was also clear that low intercorrelations might be attributed either to limited reliability of ratings or to true independence of the factors being rated, and that choice between the two explanations would depend on the degree to which sub-scores on the various orals were correlated with sub-scores on other techniques designed to measure similar components of competence. The data bearing on these two hypotheses are summarized below and discussed separately for each type of oral exercise designed to measure complex cognitive behavior.

TABLE 38

INTERCORRELATIONS OF SUB-SCORES ON PATIENT INTERVIEWS
EXAMINER I VERSUS EXAMINER II

1966 CERTIFICATION EXAMINATION

N=383

Scores Assigned By Examiner II

Scores Assigned By Examiner I

	Diagnostic Interview					Proposed Treatment Interview				Total
	1	2	3	4	5	1	2	3	4	
1. Information Gathering	.53	.37	.45	.48	.52	.39	.38	.42	.41	.52
2. Communication	.39	.36	.39	.37	.40	.39	.40	.40	.41	.46
3. Efficiency	.53	.42	.53	.52	.55	.41	.39	.42	.43	.56
4. Diagnosis	.41	.37	.43	.55	.52	.35	.37	.37	.41	.50
5. Overall	.52	.42	.50	.56	.58	.42	.43	.45	.46	.57
Proposed Treatment Interview										
1. Statements	.41	.38	.43	.43	.45	.59	.60	.62	.64	.60
2. Manner	.45	.42	.46	.46	.50	.63	.65	.65	.67	.65
3. Interaction	.41	.40	.42	.44	.47	.64	.66	.66	.69	.63
4. Overall	.45	.42	.46	.45	.50	.67	.69	.69	.72	.66
Total	.55	.48	.55	.57	.60	.60	.61	.62	.65	.69

TABLE 39: INTERCORRELATIONS OF POOLED SCORES FROM 2 EXAMINERS ON PATIENT INTERIVEWS

1966 ORTHOPEDIC CERTIFICATION EXAMINATION

N=383

	Diagnostic Interview					Proposed Treatment Interview				Total
	1	2	3	4	5	1	2	3	4	
Diagnostic Interview										
1. Information Gathering	-	.76	.83	.73	.90	.60	.57	.57	.60	.84
2. Communication	.76	-	.75	.65	.78	.63	.63	.62	.64	.83
3. Efficiency	.83	.75	-	.76	.87	.63	.58	.58	.62	.88
4. Diagnosis	.73	.65	.76	-	.90	.59	.57	.56	.59	.82
5. Overall	.90	.78	.87	.90	-	.65	.62	.62	.66	.90
Proposed Treatment Interview										
1. Statement	.60	.63	.63	.59	.65	-	.91	.91	.95	.89
2. Manner	.57	.63	.58	.57	.62	.91	-	.91	.95	.87
3. Interaction	.57	.62	.58	.56	.62	.91	.91	-	.95	.87
4. Overall	.60	.64	.62	.59	.66	.95	.95	.95	-	.90
TOTAL	.84	.76	.78	.82	.90	.89	.87	.87	.90	-

The factor structure of the sub-scores on the Diagnostic Interview was studied in the 1966 Final Certification Examination in which the Diagnostic Interview (DI) was administered as a role-playing exercise, together with the Proposed Treatment Interview (PTI). Two examiners independently scored each interview with a specially developed form, on which 3 to 4 sub-scores and an overall score were recorded separately, for each part of the interview.* The intercorrelations within and between the two raters' scores, shown in Table 38, reveal reasonably high inter-rater agreement on the Overall Score for the Proposed Treatment Interview and for the total score on the combined Diagnostic and Treatment Interviews, but substantially less agreement on the several part scores. Indeed, it is of some concern that the correlations between ratings assigned by the two examiners on different sub-scores are often as high between ratings assigned on the same sub-score. When the ratings of the two examiners are pooled, thus increasing their reliabilities, the intercorrelations between sub-scores usually exceed .70 and are, in some cases, as high as .90. (see Table 39). These data reveal a strong halo effect in the assignment of sub-scores, with the Communication Sub-Score on the Diagnostic Interview being the most independent. Subsequent correlations and factor analytic studies confirmed this impression and suggested that, in view of its relation to scores on the traditional orals and on certain components of the written examination (see Table 40), this sub-score was, in part at least, a measure of affective behavior.

TABLE 40

CORRELATION OF COMMUNICATION AND DIAGNOSIS SUB-SCORES ON THE DIAGNOSTIC INTERVIEW (DI) WITH SELECTED VARIABLES

1966 ORTHOPAEDIC CERTIFICATION EXAMINATION
N=383

	Total	Written Examinations		Oral Examinations			
Score		Multiple Choice	Simulations	Total	Pathology	Children's	Staff Conference Simulation
DI Communication	.30	.28	.11	.46	.26	.15	.18
DI Diagnosis	.20	.19	.04	.38	.20	.09	.09

---

* See Appendix 10 for a copy of the rating form and descriptions of the sub-scores.

Analyses of these data and considerations of practicality led to the decision to administer the Diagnostic Interview and Proposed Treatment Interview separately in the 1968 oral examinations and to the revision of the rating form so as to require all oral examiners to rate candidates on the same four aspects of competence: Recall, Interpretive Skill, Problem-Solving Ability and Attitudes.* Table 41 reports the intercorrelations between these sub-scores for the five oral examinations included in the 1968 Final Certification Examination. Even with the revision in procedures described above, it appears that these sub-scores are closely interrelated, that the score on the attitude component again manifests the greatest independence, and that the sub-scores on the oral simulations are generally characterized by a somewhat different pattern from that of other oral tests. While, theoretically, this could be explained as being due to differential reliabilities of the subtests, the correlations between oral examination sub-scores and other variables, including supervisors' ratings (see Table 42), suggest that the attitude score is at least as reliable as others and is substantially less influenced by the cognitive skills sampled in the multiple choice and written simulation tests. This interpretation is supported by the factor analytic studies (see Table 7) showing that the Simulation Orals do not load on the same factors as do other orals. These data suggest the advisability of collapsing the oral Recall, Problem Solving and Interpretive Skills scores. This modification was, in effect, accomplished by the weighting system that was developed and the ground rules that were established for determining the "pass-fail" levels in the 1968 Final Certification Examination.**

In summary, despite the "halo" effects in scoring oral examinations, there is considerable evidence that interpretive ability differs somewhat from what is called problem solving ability, and that both of these differ from attitudes and communication skills. Hopefully, these factors can be purified and the intercorrelations between them reduced by the development of problem exercises and rating forms that incorporate improved methods, and by the further training of examiners to utilize these techniques more reliably.

---

* See Appendix _13_ and Table 31 for a copy of the rating form and a description of the rating factors.

** See Chapter _X_ for a description of the system employed in combining sub-scores and setting standards in the 1968 Examination series.

TABLE 41

INTERCORRELATIONS OF SUBSCORES ON THE ORAL TESTS 1968 CERTIFYING EXAMINATION

N=391

Test and Rating Factor	Interpretive Skill	Problem Solving Ability	Attitudes
Problem-Solving: Adult Orthopaedics			
Recall	.83	.77	.72
Interpretive Skill		.86	.68
Problem-Solving Ability			.67
Problem-Solving: Children's Orthopaedics			
Recall	.83	.80	.77
Interpretive Skill		.85	.75
Problem-Solving Ability			.78
Problem-Solving: Trauma			
Recall	.77	.79	.60
Interpretive Skill		.86	.64
Problem-Solving Ability			.64
Observation and Interpretation			
Recall	.77	.77	.54
Interpretive Skill		.71	.54
Problem-Solving Ability			.57
Simulation Orals			
Recall	.75	.78	.71
Interpretive Skill		.83	.74
Problem-Solving Ability			.42
TOTAL: All Five Orals			
Recall	.88	.88	.81
Interpretive Skill		.90	.82
Problem-Solving Ability			.84

Concurrent Validity

The Diagnostic Interview. Given the low reliability of this exercise, as a separate test (see above), it would not be surprising to find that it lacks concurrent validity as measured by correlations with supervisory ratings. It is, therefore, of special interest to note that for third and fourth year residents the subscores on Diagnosis was the best predictor of the rating factor "Information Gathering." This result suggests that instead of rating the process of "Information Gathering," which they rarely see, chiefs of training rated the product, i.e., the accuracy of resident diagnosis. In this limited sense, the one subscore on the Diagnostic Interview may be said to have concurrent validity.

Oral Examinations of Complex Cognitive Behavior. Correlations between various criterion variables and the total scores on the 4 aspects of competence* assessed in the new types of oral tests of complex cognitive abilities incorporated in the 1968 Final Certification Examination are reported in Table 42. The low positive correlations found between subscores on the oral tests and other assessments of competence are, in part, due to the error variance in each. Despite these low reliabilities, differences in the magnitude of the several correlations (e.g., higher correlations between oral test score and supervisors' ratings of cognitive attributes than between oral test scores and their ratings of psychomotor and affective behavior) are in the expected direction. More detailed data bearing on the concurrent validity of these new orals were presented above** in the discussion of the reliability and validity of supervisors' ratings; as summarized in Table 30, these data reveal that the cognitive orals are consistently the best predictors of supervisors' ratings. It is of special importance, therefore, to note that among all the oral examinations, scores on the 30-minute Observation and Interpretation Examination have the highest partial correlations with 7 of the 10 rating factors. This phenomenon is best understood in light of the factorial complexity of that test as revealed by the factor analytic studies. Rating factors are factorially complex simply because people have difficulty abstracting the reasons why a particular individual performs the way he does. Therefore, an evaluative technique of similar complexity will show relatively high correlation with such rating factors.

---

* These aspects of competence were: Recall, Interpretive Skill, Problem-Solving Ability and Attitudes.

** See Chapter V.

TABLE 42

CORRELATIONS BETWEEN SUBSCORES ON ORAL EXAMINATIONS AND OTHER EVALUATIONS

N=391 1968 Certification Examination

Subscores on All 5 Oral Examinations	Ratings of Problem-Solving	Ratings of Surgical Technique	Supervisors Ratings Patient Relations	Supervisors Ratings Overall Competence	Scores on Written Examinations Multiple Choice Total Score	Scores on Written Examinations Simulations Total Proficiency
Recall	.30	.21	.20	.28	.47	.21
Interpretive Skill	.34	.24	.23	.30	.47	.26
Problem-Solving Ability	.31	.21	.26	.27	.47	.23
Attitudes	.28	.19	.20	.25	.36	.11

These results do not mean that the other orals are redundant or that the Observation and Interpretation format should be adopted for all of them. As revealed by the detailed data reported in Appendix 15, scores on all the oral examinations make a positive contribution to the prediction of competence. The way in which each of the orals contributes to this prediction of criterion variables, and the effects of unreliability on such predictions are shown by Table 43.

These data suggest that the cognitive oral examinations could reasonably be considered parallel versions of the same test and that the low correlations between criterion variables and subscores on individual tests are due, in part, to the unreliability of the latter. This interpretation is supported by the fact that when similar scores are combined across tests, the size of the correlation increases. The additive nature of the test data is indicated by the close similarity in the magnitude of the multiple correlations between rating variables and test scores and the simple correlations between rating factors and weighted total test scores (see Table 43.) The degree of agreement in these two sets of correlations suggests that the weighting of the tests as determined by logical criteria was quite close to the empirical weighting yielded by the multiple regression equation.

TABLE 43

CORRELATIONS BETWEEN RATING FACTORS AND OTHER TECHNIQUES
1968 FINAL CERTIFICATION EXAMINATION
SCORES

N = 391

Rating Factor	Relia-bility of Ratings	Problem Solving Orals			Sum of Prob. Solv. Orals *	Observ. and Interp Oral**	Weighted Sum of All Orals Including Simulations				Weighted Mean Total of All Tests	Mult pl R
		Adult *	Child *	Trauma *			Recall	Interp.	Prob Solv	Atti-tudes		
Info. Gathering	.24	.16	.17	.22	.30	.21	.30	.31	.28	.19	.35	.36
Prob. Solving	.29	.18	.14	.21	.29	.27	.31	.35	.28	.24	.39	.40
Clin. Judgment	.22	.16	.14	.21	.26	.22	.28	.30	.26	.21	.32	.34
Surg. Tech	.29	.14	.14	.12	.21	.18	.22	.25	.20	.15	.23	.26
Pat. Rel.	.25	.12	.09	.16	.18	.17	.20	.22	.16	.18	.21	.25
Cont. Resp.	.23	.15	.15	.11	.22	.18	.22	.25	.20	.16	.26	.29
Emer. Care	.28	.14	.13	.16	.22	.21	.26	.27	.22	.16	.26	.28
Coll. Rel.	.26	.14	.07	.10	.17	.16	.19	.21	.16	.15	.24	.26
Ethics	.28	.16	.12	.14	.21	.11	.21	.19	.21	.14	.24	.25
Overall	.31	.18	.14	.18	.27	.22	.29	.31	.26	.20	.35	.37

* Sub-score on problem-solving factor
** Sub-sc[illegible] on interpretation

In summary, the data indicate that each of the tests included in the 1968 Final Certification Examination makes an independent contribution in the prediction of competence as that is defined by supervisors' ratings, and that the factor composition of the tests is close to the factor weighting that supervisors use in evaluating their residents.

Finally, it is clear that lengthening each test from which an independent subscore was derived, increased the concurrent validity of the orals, as estimated from the correlation between test scores and criterion variables. Using the Guilford[2] technique to correct the correlation for attenuation due to unreliability, it is found that with an estimated reliability of .50 for the three problem-solving orals (probably an overestimate) the corrected correlation between the rating factor, "Problem Solving," and the Oral Problem Solving Score is .76. The square of this figure, .68, is the estimate of the proportion of true common variance between the 1-1/2 hour Problem Solving Orals and supervisors' rating of problem solving ability. A substantial amount of the remaining 32% of the variance is probably associated with the multiple choice test, the written simulations, the Observation and Interpretation Oral and the Simulation Oral. These results suggest that whatever supervisors mean when they rate problem-solving ability, is closely related to the Problem-Solving score on the three Problem-Solving Orals, since the tests reflect the same factors as observers' ratings of habitual performance.

In summary, studies of the oral examinations of complex cognitive behavior incorporated in the 1968 Final Certification Examination indicate that as presently constituted, these examinations predict about as much of the variance in ratings of habitual performance as can be expected, in view of the inherent unreliability of both ratings and oral examinations. Further improvements in these oral examinations will consist in increasing their reliability by better selection and orientation of examiners, by increasing the number of cases and/or the number of examiners, and by utilizing statistical methods to adjust for error. The results from the Orthopaedic Training Study suggest that the methods of pooling data, of structuring the examinations, of standardizing case materials and of training orals examiners described above can be employed to minimize many of the validity and reliability problems that plague traditional oral examinations and can thereby increase substantially, the arsenal of techniques available for the assessment of clinical competence.

Oral Tests of Attitudes -- The Role Playing Orals

Nature of the Examination

The critical incident study had identified the ability to communicate with and relate to patients and colleagues as among the critical components of competence. As noted above, two role playing exercises, the Simulated Diagnostic Interview and the Proposed Treatment Interview, were the first examinations specifically developed to assess these aspects of performance. Subsequent to their experimental introduction on the 1966 oral examinations the Simulated Diagnostic Interview was converted from a role-playing to problem-solving exercise and was incorporated in the Problem-Solving Orals discussed above; the Proposed Treatment Interview was maintained as a role-playing exercise and administered together with two additional types of simulated confrontations in a separate half-hour examination. These additional simulations included such situations as discussing guarded prognosis with a patient, instructing a nurse to modify her handling of a patient, explaining the failure of a treatment to a patient, conferring with a colleague with whom one disagrees, and the like. In each such exercise, the candidate was given about 3 minutes to read a brief description of the situation, such as that shown in Exhibit V, opposite; he was then given approximately 7 minutes to talk with the simulated patient or colleague, whose role was taken by an examiner. The candidate's performance was scored independently by two examiners on a standardized rating form (see Appendix 10). Data on the reliability and validity of the Simulation Orals are reported below.

Reliability of Simulation Orals

As with the oral examinations of complex cognitive skills, there are two possible sources of unreliability in the scoring of oral tests of attitudes: inter-rater disagreements and sampling errors. Data on the former, summarized from several studies in Table 44, indicate that there is a relatively high level of agreement in the independent ratings assigned by two examiners observing the same examination and that the inter-rater reliability of the examination was not significantly altered by a three-fold increase in its length. This somewhat surprising finding may be accounted for by the fact that in the 1966-67 studies, the 10-minute Proposed Treatment Interview was administered together with a 20-minute Simulated Diagnostic Interview, and examiners may have been strongly influenced by performance on the latter in their ratings of the former, thus, in effect, basing their judgment on a half-hour sample of candidate behavior.

EXHIBIT V : Candidate Instructions for Sample Proposed Treatment Interview

As the orthopedist-on-call for the day, you are called to the emergency room of your hospital to see a patient whom you have never met before. The patient is a 28-year old male laborer who has a two day history of severe low back pain which radiates down the left posterior thigh into the calf; there is also some tingling and numbness over the left lateral calf. The pain began when he injured his back while lifting a heavy weight from the floor; he has had no previous episode of this type. The pain is aggravated by coughing.

On examination there is marked paravertebral muscle spasm; the patient is listing to the right. There is marked tenderness at the L4-L5 interspace. Back motion is restricted in all planes. Straight leg raising on the left produces pain in the left leg at 30° elevation; straight leg raising on the right produces pain in the left leg at 60°. There are no reflex changes, no motor weaknesses, and no sensory changes. Otherwise, the physical examination is within normal limits.

You have reviewed this patient's X-rays, one of which is enclosed.

You suggest to the patient that he enter the hospital for a period of complete rest and "conservative care." Your suggestion of hospitalization immediately alarms the patient, who then asks, "What _is_ wrong with my back, Doctor?"

You will now describe to a simulated patient what is wrong with "his" back and explain why you proposed your course of treatment. The laborer's name is Mark Cole.

TABLE 44

RELIABILITY OF ORAL SIMULATIONS BASED ON SEVERAL STUDIES

Technique	Length of Exercise	N	Examination	Mean	Correlation Between Raters Observing Same Examination	Overall Reliability
Proposed Treatment Interview	7-10 min.	383	1966 Orthopaedic Certification Examination	7.1	.72	.84
Proposed Treatment Interview	7-10 min.	30	1966 Orthopaedic Certification Examination	6.9	.55	.71
Proposed Treatment Interview	7-10 min.	387	1967 Orthopaedic Certification Examination	7.0	.61	.76
Simulated Interview	28-30 min.	391	1968 Orthopaedic Certification Examination	7.7	.73	.84

TABLE 45

MEAN SCORES OF RESIDENTS AT DIFFERENT LEVELS OF TRAINING ON THE PROPOSED TREATMENT INTERVIEW

1966 Oral In-Training Examinination

Level	N	Mean*	S. D.	
1st Year	29	6.2	2.8	
2nd Year	75	6.9	2.8	
3rd Year	50	6.5	2.8	
4th Year	75	7.5	2.6	
Total	233	6.9	2.6	

* .ANOVA indicates a P. Value 4 .08.

In the one limited study of the combined effects of rater and sampling error on the reliability of the simulation orals, two different Proposed Treatment Interviews were separately administered by two examiners to a group of 25 residents in the 1966 In-Training Examination. The correlation between the scores on the two examinations was .49; combining the two scores yields an estimated reliability of .72. This is an astonishingly high value for a 7-10 minute test and is of special interest in view of the low correlation across cases in the problem solving orals and the written simulation exercises. The higher sampling reliability of the Simulation Orals is probably attributable to the fact that they presuppose much less specific content than do the Problem Solving Orals and written simulations.

Unfortunately, it was impossible; for mechanical reasons, to study either the rater or sampling error of the Simulation Orals included in the 1968 Final Certification Examination. However, the two studies reported above strongly suggest that those orals consisting as they did of a half-hour session devoted to 3 or more standardized exercises administered and rated by two examiners reached an acceptable level of reliability.

Validity of Simulation Orals

Content validity of the simulation orals was assessed by both observational analysis and systematic questionnaire studies. The former, discussed above,* indicates that the behavior required of examinees in the Simulation Orals was quite unlike that demanded of them in either the traditional orals or the new Problem Solving Orals. Furthermore, the observed differences in the nature of examiner-candidate exchanges in the 3 types of oral formats were in the hypothesized direction. The results from the questionnaire study** were similarly encouraging. Specifically, some 80% of the examiners were convinced that the simulations provided valuable information concerning the candidate's ability. Approximately two-thirds of the candidates agreed that the examination gave them a chance to demonstrate their ability in some important area of orthopaedic surgery, and almost three-fourths felt that most of the topics covered were important to orthopaedic practice; only 12% reported that the examination procedures were confusing, despite the fact that role-playing was new to most. In short, the role-playing simulations met acceptable standards of content validity as samples of specified attitudes and skills in patient and colleague relations.

---

* See Table 34 and Appendix 28 for additional data on the observational study.

** See Tables 35 and 36, and Appendices 18 and 19 for detailed tabulation of examiner and candidate responses in the questionnaire study.

Construct Validity of the Simulation Orals was investigated in two separate studies. In the first, conducted in connection with the 1966 In-Training Examination, the relationship between level of training and scores on the Proposed Treatment Interview was analyzed. The results, summarized in Table 45, indicate slight, though not statistically significant, improvement in performance. Differences in scores between groups at different levels of training were substantially greater on both the Diagnostic Interview and the conventional orals. While this finding may appear to raise doubt about the construct validity of the Simulation Orals, it should be considered in light of the fact that most supervisors of orthopaedic training programs report that they almost never observe their residents dealing with patients and that little attention is paid to their ability to relate to, and communicate with, patients. For example, in the questionnaire study referred to above, * some 40% of the examiners administering Simulation Orals agreed with the statement: "Most training programs apparently do not adequately train the candidates to take this type of examination." In contrast, only 27% of those administering the Observation and Interpretation Orals, and fewer than 20% of those administering the Problem-Solving Orals, agreed with the statement quoted. In short, the findings regarding the relationship between level of training and performance on the Simulation Orals are compatible with supervisors' expectations as based on the nature of the training programs.

The second study of the construct validity of the Simulation Orals, conducted in connection with the 1968 Final Certification Examination, consisted in a correlational analysis of the interrelations among sub-scores on that oral, and between it and supervisors' ratings of habitual performance. The results, summarized in Tables46 and 47, reveal that the intercorrelations of different types of sub-scores on the Simulation Orals are substantially higher than those between similar sub-scores on different types of examinations, and higher than those between the various sub-scores and the relevant rating factors. In short, the data indicate that the cognitive and attitudinal sub-scores on the Simulation Orals are not independent.

---

* See Tables 35 and 36 and Appendices 18 and 19 for detailed tabulation of examiner and candidate responses in the questionnaire study.

TABLE 46

CORRELATIONS BETWEEN SUB-SCORES ON SIMULATION ORALS AND SELECTED OTHER EVALUATION TECHNIQUES

1968 Final Certification Examination

Scores on Selected Variables	Sub-Scores on Simulation Orals			
	Recall	Interpretation	Problem-Solving	Attitudes
Rating Form - Problem Solving	.15	.15	.16	.19
Rating Form - Patient Relations	.13	.12	.14	.13
Rating Form - Overall Competence	.13	.11	.11	.15
Multiple Choice - Recall	.13	.14	.14	.11
Multiple Choice - Problem Solving	.14	.13	.12	.14
Sum of Oral Problem Solving Tests - Problem Solving	.23	.25	.25	.31
Sum of Oral Problem Solving Tests - Attitude	.26	.23	.26	.31
Oral Observation & Interpretation - Interpretation	.17	.17	.19	.22
Oral Observation & Interpretation - Attitude	.21	.16	.18	.20
Written Simulation Exercises - Diagnostic Proficiency	.03	.08	.11	.02
Written Simulation Exercises - Treatment Proficiency	.04	.12	.07	.07
Written Simulation Exercises - Total	.04	.11	.11	.04
Simulation Oral - Recall	-	.77	.77	.71
Simulation Oral - Interpretation	.77	-	.83	.74
Simulation Oral - Problem Solving	.77	.83	-	.82
Simulation Oral - Attitudes	.71	.74	.82	-

TABLE 47

COMPARISON OF ATTITUDE SCORES ON ORALS WITH SELECTED RATING FACTORS AND PROBLEM SOLVING EXERCISES

1968 ORTHOPEDIC CERTIFICATION EXAMINATION

N=391

ORAL TEST SCORES	RATING FACTORS						TOTAL TEST SCORES	
	INFORMATION GATHERING	PROBLEM SOLVING	PATIENT RELATIONS	COLLEAGUE RELATIONS	ETHICS	OVERALL COMPETENCE	MULTIPLE CHOICE	WRITTEN SIMULATIONS
Simulation-Examiner-I	.13	.17	.13	.08	.08	.12	.20	.02
Simulation-Examiner-II	.14	.18	.16	.14	.14	.15	.15	.05
Simulation-Sum of Scores Examiners I and II	.14	.19	.17	.12	.12	.15	.19	.04
Adult Problems-Problem Solving	.16	.18	.12	.14	.16	.18	.35	.11
Children's Problems Problem Solving	.17	.14	.09	.07	.12	.14	.23	.13
Trauma Problems-Problem Solving	.22	.21	.10	.10	.14	.18	.34	.17
Observation and Interpretation-Interpretation	.21	.27	.17	.16	.11	.22	.29	.22

Second, they suggest that in addition to the general factor of "handling oneself in encounters with others" (a factor which must saturate not only all the oral examinations, but the rating data as well), the Simulation Orals measure some aspect of competence different from that assayed by the Problem-Solving Orals. This interpretation is based on the hypothesis that if the two types of examinations were measuring the same types of competence, scores on the Simulation Orals would be more highly correlated with other test scores due simply to their higher reliability. Such is not the case; scores on the Simulation Orals show significantly lower correlations with scores on multiple choice and written simulation exercises than do scores on the other oral examinations. While this finding probably reflects the fact that less specific content is required in responding to the Simulation Oral than to other orals; these data do not, in themselves, indicate whether the other components of competence measured in the Simulation Oral are, in fact, skill and ability in relating to, and communicating, with patients and colleagues. In short, the data are compatible with the assumption of reasonable construct validity of the Simulation Orals, but are by no means conclusive in establishing it.

Concurrent Validity of the Simulation Orals was investigated in two studies of the relationship between scores on that examination and supervisors' ratings of various aspects of trainees' habitual performance. Though this method was employed in the estimation of the concurrent validity of all assessment techniques developed in the study, its application in the evaluation of the Simulation Orals presented certain special difficulties that should be noted. Specifically, in addition to the inherent unreliability of both the rating data and the scores on oral tests, the rating data, in this case, lacked validity for the following reasons:

Supervisors rarely observe residents with patients; hence to the extent that rating data are distorted by "halo" effects, ratings will be strongly influenced by what supervisors regard as problem-solving ability rather than by a general factor of ability to relate to patients and colleagues. Similarly, because of limited insight and experience in analyzing the dynamics of interpersonal relationships, some examiners have difficulty rating these factors in the oral tests and thus, given the "halo" effect, tend to identify communicative ability with problem solving ability. Finally, because of the artificiality of the situation and the constaints placed on it by the presence of examiners, examinees may feel that they have little opportunity to demonstrate their ability to relate to, and communicate with, patients and colleagues. It is probable that all of these factors play a part in depressing the correlation between scores on Simulation Orals and relevant criterion factors in supervisors' ratings. However, despite their handicaps the Simulation Orals do make an important contribution to the prediction of supervisors' ratings of ability to relate to patients; additionally, they contribute about as much to the prediction of other criterion factors as do the other half-hour orals. Finally, the data suggest that if such orals are to be included as a regular part of the certification examination, major effort should be devoted to improvements in the scoring system used and to the development of techniques that would yield better criterion data for affective variables.

Oral Tests of Attitudes and Skills--The Simulated Staff Conference

Nature of the Examination

One additional oral technique, incorporated experimentally in the 1966 and 1967 Final Certification Examination consisted in a simulated staff conference in which five candidates discussed one or two cases. Candidates were given five minutes to study the protocol of the cases and 22 minutes to discuss them:* One or more examiners observed and rated the entire proceedings, without commenting on them or participating directly in any way. Two different rating forms were used for scoring these conferences: In the first, adapted from the work of Bass,[3] the candidate was rated on four factors: Individual achievement, Ability to assist the group to reach its goals, Effective conduct as member of a group, and Overall competence.** In the second scoring form, subsequently adapted from Bales[4] technique, the examiner was directed to classify and tally each statement by each participant on the following scale:

0 Error (in content)
1 Hinders group
2 Is passive non-facilitative
3 Clarifies and provides constructive suggestions
4 Organizes, integrates, greatly facilitates.

On the basis of the quality and quantity of the contributions tallied for each candidate, the examiner was expected to rate his overall competence and to record this judgment on the familiar 12-point scale.*** Studies of the reliability and validity of this technique are summarized below.

---

* See Appendix 20 for a typical case protocol.
** See Appendix 12 for a copy of this form.
*** See Appendix 11 for a copy of this form.

Reliability of the Simulated Staff Conference

No study of intercase reliability was conducted; in the one study of interrater agreement the correlation between the scores assigned by two examiners observing the same conference was found to be .14 for the 383 candidates rated in the 1966 Final Certification Examination; this would yield an estimated reliability of .25 for the pooled scores of the two examiners. The results for each of the three examining teams involved in that examination are summarized in Table 48. These data indicate that though the level of agreement was not high for any team, one set of examiners was apparently in complete disagreement on standards.

For this reason, modifications were made in the scoring technique as described above and the examination was repeated in the 1967 Final Certification Examination. However, the Board was able to assign only one examiner to administer and score each oral; it was, therefore, impossible to carry out any reliability studies on the new scoring methods.

Validity of the Simulated Staff Conference

Content Validity of this oral technique was assessed on the basis of reports from candidates and examiners, many of whom agreed that the simulated discussions preserved much of the "feel" of staff conferences commonly held during residency. Some have criticized the technique as being too artificial and predicted that no candidate, however rude and tactless in a "real situation," would display these characteristics in the simulated situation, a fear that has not been justified. Finally, some have criticized the technique as sampling situations which, though common during residency training, are not characteristic of practice.

Construct Validity of the Simulated Staff Conference was studied in the 1966 Final Certification Examination. The data, summarized in Table 49, reveal that, despite low interrater agreement in scoring these exercises, the pattern of correlation between scores on that test and scores on other types of tests does not differ significantly from that characteristic of other simulation techniques (the Diagnostic Interview and Proposed Treatment Interview). Similarly, the factor analytic studies of the 1966 Final Certification Examination suggest that the factor structure of the Simulated Staff Conference does not differ significantly from that of the more conventional orals.

TABLE 48

CORRELATION BETWEEN SCORES ON OVERALL COMPETENCE IN SIMULATED STAFF CONFERENCES

1966 Final Certification Examination

Team	N	Mean Score	Correlations Between Scores of Two Examiners
1	144	8.2	.48
2	123	7.6	.30
3	116	8.1	-.49
TOTAL	383	8.0	.14

TABLE 49

CORRELATION OF SCORES ON SIMULATION ORALS WITH SELECTED OTHER TECHNIQUES

1966 Final Certification Examination

N = 383

					Traditional Oral Examinations				
ORAL TEST	GRAND TOTAL	MULTIPLE CHOICE	SHORT ANSWER	TOTAL ORAL	PATHOLOGY	CHILDREN'S ORTHOPAEDIC	ANATOMY TRAUMA	ADULT	WRITTEN SIMULATIONS
Simulated Staff Conference*	.23 .23	.23 .23	.22 .22	.30 .30	.24 .24	.18 .18	.13 .13	.20 .20	.06 .06
Diagnostic Interview	.25	.23	.25	.47	.27	.16	.17	.28	.06
Proposed treatment Interview	.25	.23	.22	.57	.28	.19	.19	.33	.09

* Not included in Grand Total and Total Oral Scores

Discontinuation of the Simulated Staff Conference

Because of its low reliability as initially scored, and its dubious validity (see above) the Board decided not to continue the Simulated Staff Conference as a regular part of the certification process. In this discussion, they were supported by the following reservations shared by numerous examiners:

(1) Some were not convinced of the importance of the behavior the simulation was designed to assess.

(2) Some were skeptical about making a decision about a candidate on the basis of his 5 or 6 minute participation in a group discussion; and many felt that the constitution of the group affected each person's performance and thus feared that an individual might look bad because of the group he was in and not because of any real weakness on his part.

In short the Simulated Patient Management Conference proved to be an interesting technique that had to be abandoned due to inadequate rater reliability and to lack of acceptance from the profession; nevertheless, work done by Bass[3] and others indicates that the technique may have great usefulness which has not, as yet, been exploited by the orthopaedic profession.

Summary

The assessment of each new type of oral examination recommended for incorporation in the certification examinations, included considerations of both its reliability and its validity. With respect to the former, every effort was made to estimate error variance due both to inter-rater disagreements and to sampling errors. Except for the simulated staff conference, all types of exercises met adequate standards of interrater reliability. Greater difficulties arose in regard to inter-case reliability, particularly in the Problem-Solving Orals in which command of specific content was more critical than in the Simulation Orals. With respect to validity, every effort was made to investigate the content, construct and concurrent validity of the new techniques. Observational and systematic questionnaire studies indicated that, with the possible exception of the Simulated Staff Conference, all techniques met adequate criteria of content validity. Correlational and factor analytic studies of construct validity suggested that, in general, scores on the new techniques were associated with each other and with scores on more conventional examinations, in the hypothesized manner. Factor analytic and multiple regression studies indicated that, in general, scores on the new techniques were associated with criterion variables in the hypothesized directions; however, due to both sampling and rater error in the measurement of both test and criterion variables, the predicted association was in several cases, weaker than anticipated.

On the basis of these data, the American Board of Orthopaedic Surgery has reconstituted its oral examinations to include 1 1/2 hours of problem-solving exercises with each half-hour administered by a different examiner, a half-hour of observation and interpretation exercises administered by a fourth examiner, and a half-hour of simulated patient and colleague confrontations administered by a team of 2 examiners. Standardized case materials and standardized scoring forms are used for all exercises. The same four factors, (recall, interpretive skill, problem-solving ability and attitudes and communication skills), are rated in each set of exercises; differentially weighted scores on each factor are pooled across all examinations. These innovations in the nature of the oral examinations and the method of scoring them are direct outcomes of the study to date.

1. Ebel, Robert, "Estimation of the Reliability of Ratings" in Principles of Educational and Psychological Measurements, ed. by Mehrens & Ebel (Rand McNally Co., 1967)

2. Guilford, J.P. "Fundamental Statistics in Psychology and Education." McGraw-Hill Book Co., Inc. New York 1956

3. Bass, Bernard, "The Leaderless Group Discussion," Psychological Bulletin, 1954 Vol. 51 pp. 465-491.

4. Bales, R. Freed, Interaction Process Analysis, Cambridge: Addison - Wesley, 1951

CHAPTER VIII:

THE MULTIPLE CHOICE QUESTIONS

The Multiple Choice technique was discussed in some detail above in connection with the analysis of previous evaluation techniques.* While many of the multiple choice questions used in 1964 were still relevant in 1968, numerous modifications occurred in the interim, with respect to procedures for developing, reviewing and scoring such exercises. As noted above, the two serious weaknesses in the multiple choice technique revealed by the earlier analysis were: (1) the tendency in these exercises to focus on measuring the recall of isolated bits of information, that often appeared to have little relation to any behavior required in the practice of orthopaedics; and (2) the tendency to set the passing mark on the basis of the distribution of scores, rather than on the basis of pre-determined standards, a tendency which, by punishing an arbitrary number of examinees for scoring at the bottom of the distribution, violates the mission of any Board established to determine whether an individual meets professional standards of competence.

To alleviate some of these problems and to encourage question authors to submit case-oriented materials that demand application of principles rather than simple recall of isolated facts, the Board developed an item classification guide with instructions that each question submitted for the certifying examination must be suitable for meaningful classification with respect to at least four of the following five dimensions:**

I. Type of patient (adult or child)

II. Type of disorder (trauma or disease, etc.)

III. Part of Body (upper extremity, etc...)

IV. Basic science (anatomy, etc...)

V. Clinical (diagnosis, etc.)

Second, utilizing these dimensions, together with a taxonomy of intellectual processes, the Board constructed a blueprint*** for the overall examination which stipulates the proportion of the total examination to be devoted to each type of intellectual process and to each content category within each of the five dimensions listed above. Third, the Board established a Task Force on New Multiple Choice Questions* and gave it the clear charge to develop a system for constructing and submitting questions that would assure a larger proportion at higher taxonomic levels.

---

* See Chapter III.

** See Appendix 21 for a copy of the most recent Item Classification Guide.

*** See Appendix 22 for a copy of the total blueprint.

With the development of the item classification guide and the blueprint, it is now possible to direct staff to pull questions of known characteristics from the item pool to fit the specifications dictated by the blueprint for a given examination. This initial draft, approximately 50% longer than will ultimately be required, is circulated to the entire Board. Each member of the Board is requested to respond to each question, to classify it according to the intellectual process it samples and to comment on its merits. The examination committee of the Board then meets, reviews the responses of the Board and makes the final selection of questions for the examination.

Finally, prior to the administration of the test, the Board establishes standards of minimum satisfactory performance.* For the multiple choice questions this entails an adaptation of the Nedelsky[1] technique, in which members of the Examination Committee review each question and check each option that a barely passing candidate should be able to eliminate. The reciprocal of the remaining number of options is taken as the "minimum passing level" (MPL), for that question. Thus, for example, if a question has five options and two are eliminated, then the chances that a minimally competent candidate would get the right answer by guessing, would be one in three or .33. If the whole test consists of such questions, the barely passing candidate should score at least 33%. The average of the MPL's for all of the questions in an examination is the best estimate of the score that a candidate would get if he eliminated all of the alternatives that informed judges think he should be able to exclude and selected among the others by guessing. Employing this technique, it is possible to transform the scores to any scale for combination with other test data. In the present study, all scores and sub-scores for both oral and written tests were converted to a 12-point scale in which the MPL was always defined as 3.5.

Application of these pre-determined standards in the 1968 Final Certification Examination resulted in a failure rate on the Multiple Choice Recall Sub-Score of approximately 50% even among graduates of American Medical Schools who were taking the examination for the first time. However, when the sub-scores on the Multiple Choice component were combined with data about performance on other types of exercises, the failure rate fell substantially.*

---

* See Appendix 26 for a detailed description of the system of setting absolute standards.

Data on the reliability and validity of the newly revised Multiple Choice examination are reported in the following sections.

Reliability of the Multiple Choice Examination

The great strength of the multiple choice technique is its consistently high reliability: In tests which are carefully constructed, and are composed of questions that have been widely reviewed, rating errors are minimized, and sampling reliability is assured by the fact that in the typical examination, it is possible to use large numbers of completely independent items. Table 50, which summarizes the reliability data on the multiple choice examinations used by the orthopaedic profession, over the last few years, reveals that the estimated reliability of these examinations varies directly with the number of items and indicates that sub-scores, based on relatively few items, are not sufficiently reliable to use independently in the certification process.

Validity of the Multiple Choice Examination

Content Validity

The Task Force on New Multiple Choice Questions was given the charge to develop questions at higher taxonomic levels for the Board examinations; the methods it established for constructing and reviewing new questions has resulted in an increased proportion of items being rated by authors and reviewers, as sampling higher cognitive processes. However, despite these improved procedures, data from the questionnaire study of examiner and candidate reaction to the new examinations (see Table 51 and Appendix 19) indicate that candidates, successful and unsuccessful alike, found the multiple choice component least relevant and least appropriate. They were especially critical of many questions which seemed to them to be ambiguous or to demand information that is remote to their needs as practitioners. Some support for their views is to be found in the fact that the relatively reliable Multiple Choice Recall test is less useful than the relatively unreliable orals as a predictor of such criteria as supervisors' ratings.

---

* See Chapter X, below for a detailed account of the scoring procedures and results of the 1968 Certification Examination.

Table 50

RELIABILITY OF MULTIPLE CHOICE TESTS

Examination	Score	Number of Items	Reliability*
1966 Final Certification Examination	Multiple Choice Total	150	.72
1966 Final Certification Examination I	Multiple Choice Total	230	.89
1967 Final Certification Examination	Multiple Choice Total	177	.84
1968 Final Certification Examination	Multiple Choice Total	180	.83
	Multiple Choice Recall	135	.71
	Multiple Choice Problem-Solving	34	.30

* Estimated by Kuder Richardson Formula 20.

TABLE 51

CANDIDATE REACTION TO COMPONENTS OF THE 1968 CERTIFYING EXAMINATION

		Per Cent Agreement with Statement as Applied To:						
		Written Examinations		Oral Examinations				
Statement	Group	Multiple Choice	Simu-lations	Problem Solving Trauma	Problem Solving Adult	Problem Solving Children's	Observation and Interpretation	Simulation
Gave me a chance to demonstrate my abilities in some important areas of orthopaedic surgery	100 who passed	43	61	86	75	82	69	67
	100 who failed	42	52	85	70	75	65	65
Most topics covered were irrelevant to orthopaedic practice	100 who passed	29	3	3	4	5	11	9
	100 who failed	20	4	4	7	3	8	3
Examination was too difficult	100 who passed	33	8	3	4	4	5	6
	100 who failed	36	10	3	7	5	17	6

TABLE 52

SCORES ON MULTIPLE CHOICE TESTS AND SUBTESTS ANALYZED BY YEAR OF TRAINING

1967 In-Training Examination

N=1682

Subtest	Year of Training									
	First		Second		Third		Fourth		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
General Orthopaedics	49.4	8.5	54.0	9.6	58.6	10.0	61.8	10.7	56.6	9.8
Adult Orthopaedics	43.7	7.5	47.7	8.4	52.1	8.8	55.5	8.7	50.3	8.5
Child Orthopaedics	49.2	9.7	54.8	9.6	61.1	9.7	65.1	9.0	58.3	9.5
Trauma	50.6	9.8	55.6	8.9	60.3	9.3	63.2	8.8	58.1	9.1
Hand Surgery	44.3	11.5	49.7	11.3	56.3	11.3	60.6	10.9	53.5	11.2
Anatomy	43.1	10.5	48.3	10.4	55.2	11.2	60.4	11.4	52.5	10.9
Pathology	46.0	9.3	50.4	9.9	56.2	9.5	60.0	9.6	53.8	9.6
Physiology and Biochemistry	51.8	8.0	55.2	8.2	56.4	8.3	57.3	8.6	55.6	8.3
Biochemistry Mechanics	40.7	0.7	46.4	11.2	52.6	11.4	57.1	11.3	50.0	11.2
Rehabilitation	37.6	13.2	43.8	14.6	51.7	15.7	56.3	14.0	48.3	14.5
Total	47.8	6.9	52.6	7.3	57.7	7.6	61.1	7.4	55.5	7.3

Construct Validity

Given the candidates' criticism of the multiple choice examination, and its lack of association with supervisors' ratings, it is always a little surprising to find, as revealed in Table 52, that there is a consistent growth, over the four years of orthopaedic residency training, in the abilities measured by the Multiple Choice questions. While differences in mean scores for groups at different levels of training are in the expected direction on all sub-tests, the amount of growth is far from uniform in the various disciplines.

TABLE 53

DIFFERENCES IN MEAN SCORES OF FIRST AND FOURTH YEAR RESIDENTS
1967 IN TRAINING EXAMINATION

N=1682

Subtest	Difference in %	Difference Divided by Standard Deviation
General Orthopaedics	12.4	1.3
Adult Orthopaedics	11.8	1.4
Children's Orthopaedics	15.9	1.7
Trauma	12.6	1.4
Hand Surgery	16.3	1.5
Anatomy	17.3	1.6
Pathology	14.0	1.5
Physiology and Biochemistry	5.5	0.7
Biomechanics	16.4	1.5
Rehabilitation	18.7	1.3
Total	13.3	1.8

It is interesting to note (see Table 53) that much greater improvement occurs in scores on Pathology and Anatomy disciplines which are directly applicable to clinical problems in surgical specialties than in scores on Physiology and Biochemistry. Such results suggest either that the latter content areas are less effectively handled in the training programs or alternately, that the examination questions in these disciplines are not probing important areas of competence. Perhaps both hypotheses have an element of truth. In an extensive review of the individual questions that discriminated most between first and fourth year residents, Dr. Huncke[2] observed that the most discriminating questions were those involving complex multi-system widespread entities, such as meningomyelocele, cerebral palsy and scoliosis. He adds: "This would seem logical since these are complicated situations that do require a considerable amount of training and understanding to proceed with any amount of accuracy. It should be noted, however, that there were questions involving these entities that, in my opinion, could have been adequately answered if the individual correlated basic science knowledge which he is presumed to have, particularly his knowledge of functional anatomy. Judging from the results, however, this was not often done by the candidates, who seemed to approach these particular complex problems as if they demanded recall rather than application."

Data on a second aspect of construct validity, i.e., the relationship between performance on the Multiple Choice Examination and on other other evaluative techniques, are summarized in Tables 54 and 55. These data clearly indicate that sub-scores on the Multiple Choice examination behave in the expected manner in relation to other test scores, and strongly suggest that the Multiple Choice test measures, primarily, cognitive functioning. For example, the score on Multiple Choice-Recall has a very low correlation with the rating factor, "Patient Relationships", and with scores on the Simulation Orals. As compared with the Recall score, that on Multiple Choice-Problem Solving, when corrected for attenuation, has a generally higher correlation with all types of assessment other than the rating factor, "Surgical Technique", and certain of the Written Simulation Scores. These exceptions may be due either to random errors of measurement, or to the fact that both the rating factor and the written simulations include important non-cognitive factors of temperament and skill which are not sampled by multiple choice techniques. Further, it is of interest to note that when corrections are made for unreliability the correlation between the Multiple Choice-Recall and the Multiple Choice-Problem Solving is .89; i.e., about 80% of the variance in the two tests is common.

Finally, additional data on the construct validity of the multiple choice technique are furnished in the factor analyses of the 1966 In-Training Examination and the 1968 Final Certifying Examination (see Tables 6, 7, and 56) both of which indicate that the multiple choice scores load heavily on one factor which appears to be a content or information factor; in contrast, other techniques show a much more complex factor structure.

TABLE 54

CORRELATIONS BETWEEN MULTIPLE CHOICE SUBSCORES AND OTHER SELECTED VARIABLES

1968 Final Certification Examination

N=391

Other Variables	Multiple Choice Recall		Multiple Choice Problem Solving	
	Actual	Corrected *	Actual	Corrected*
Rating Factors				
Information Gathering	.25 21	.30	.20 .19	.37
Problem Solving	.31 28	.37	.26 .23	.48
Clinical Judgment	.23 24	.27	.19	.35
Surgical Technique	.16 13	.19	.10 .12	.19
Patient Relationships	.09 10	.11	.08 .09	.15
Cont. Resp.	.16	.19	.12 11	.22
Emergency Care	.17 15	.20	.15 14	.28
Overall Relationships	.15 14	.18	.12	.22
Ethics	.19 16	.23	.12 13	.22
Overall Competence	.28 26	.33	.24 .22	.44
Oral Tests				
Adult	.27	.32	.20	.37
Child	.15	.18	.07	.13
Trauma	.30	.36	.23	.43
Interpretive	.22	.26	.21	.39
Simulation Attitudes	.12	.14	.14	.26
Written Simulation				
Diagnosis Proficiency	.21	.25	.17	.31
Treatment Proficiency	.19	.23	.11	.20
Total Proficiency	.26	.31	.17	.31
Multiple Choice				
Recall Actual	—	—	.41	.76
Recall Corrected	—	—	.49	.89

* Corrected by Guilford technique for attenuation, due to unreliability of the Multiple Choice Sub-Tests.

TABLE 55

CORRELATIONS BETWEEN MULTIPLE CHOICE SCORES AND OTHER SELECTED VARIABLES BY YEAR OF TRAINING

1966 In-Training Examination

Other Variables	First and Second Years N=109 Actual	Third and Fourth Years N=119 Actual	Total All Four Years N=228	
			Actual	Corrected for Unreliability on MC Section
Rating Factors				
Recall	.25	.29	.31	.33
Problem-Solving	.23	.25	.26	.28
Information-Gathering	.19	.43	.33	.35
Clinical Judgement	.23	.24	.26	.28
Patient Relations	.13	.09	.19	.20
Colleague Relations	.09	.09	.07	.07
Surgical Skill	.13	.26	.14	.15
Ethics	.17	.29	.18	.19
Overall	.20	.26	.26	.28
Other Test Scores:				
Diagnostic Interview	.17	.20	.26	.28
Proposed Treatment Interview	.28	.23	.27	.29
Traditional Adult Oral	.23	.36	.44	.47
Written Simulations Total Proficiency	.05	.18	.01	.01

Concurrent Validity

The two major studies of the concurrent validity of the multiple choice technique were conducted on the 1966 In-Training and the 1968 Certification Examinations. Despite the fact that the correlations reported in Tables 54 and 55, between scores on the multiple choice test and supervisors' ratings or other test scores are generally rather low, differences among the several values are in the expected direction. Specifically, the same general pattern of relationships characterizes both the 1968 certification and the 1966 In-Training data; this general pattern is one, in which, despite validity and reliability problems in the ratings that tend to depress all correlations in the matrix the multiple choice scores are significantly more closely related to ratings of cognitive components of competence than to ratings of skills and affect. Similar patterns characterize the relationships between scores on multiple choice tests and scores on other written and oral tests.

These data indicate that the pattern of test score predictors shifts from one rating factor to another despite the fact that the intercorrelations among ratings on the several factors is quite high.

TABLE 56

RESULTS OF MULTIPLE CORRELATIONAL ANALYSIS
1968 CERTIFYING EXAMINATION

N=391

Rating Factors as Dependent Variables	Multiple		Test Scores as Independent Variables	Partial	
	R	F		r	F
Information Gathering	.36	5.13	Multiple Choice Recall	.12	5.56*
			Observation and Interpretation Interpretation	.11	4.48*
			Trauma-Problem Solving	.10	3.84
Overall Competence	.36	5.13	Observation and Interpretation Interpretation	.12	5.48*
			Multiple Choice Recall	.12	5.46*
			Multiple Choice Problem Solving	.09	3.13

* Significant at .05 level of confidence
** Significant at .01 level of confidence

As regards the concurrent validity of the Multiple Choice Test, it is important to note that it is the best predictor of ratings on "Information Gathering," the second best for ratings of "Problem Solving," the third best for ratings on "Clinical Judgment" and disappears as an important predictor for rating factors related to affective behavior, i.e., "Patient Relationships," and "Colleague Relationships." (Table 56).

Secondly, when appropriate adjustments are made for differences in the reliabilities of candidate and resident ratings, it appears that the Certification Examination identifies a much larger proportion of the true variance in the criterion data than does the In-Training Examination (Table 57).

TABLE 57

COMPARISON OF THE PREDICTORS OF OVERALL COMPETENCE
1968 FINAL CERTIFYING EXAMINATION AND
1966 IN-TRAINING EXAMINATION

Examination	Reliability of Rating Factor	Multiple R (All Test Scores)	Multiple R Corrected for Unreliability of Rating Factor
1966 In-Training Examination	.73	.32	.38
1968 Final Certifying Examination	.31	.36	.65

This is probably attributable, in part, to improvements between 1966 and 1968 in the Multiple Choice section of the examination and specifically to the development of the problem solving subtest in the Multiple Choice format, as well as to revisions and extensions of the oral test techniques.

Summary Comment

In summary, the multiple choice technique, as modified in the current study, provides valuable information on certain facets of competence in orthopaedic surgery, as these are defined by supervisors' ratings. However, it is necessary to supplement this method of assessment with other techniques, in order to obtain valid and reliable data on all aspects of competence in the specialty. Such supplementation is of special value in evaluating those areas of competence that involve affective behavior.

1 Nedelsky, Leo, "Absolute Grading Standars for Objective Tests", Educational and Psychological Measurement, Vol. 14, Spring, 1954 pp. 3-19.

2 Huncke, Brian H., M.D., Memorandum, "Review of Discriminating Questions", in the November, 1967 In-Training Examination, May 14, 1968. N.P.

CHAPTER IX:

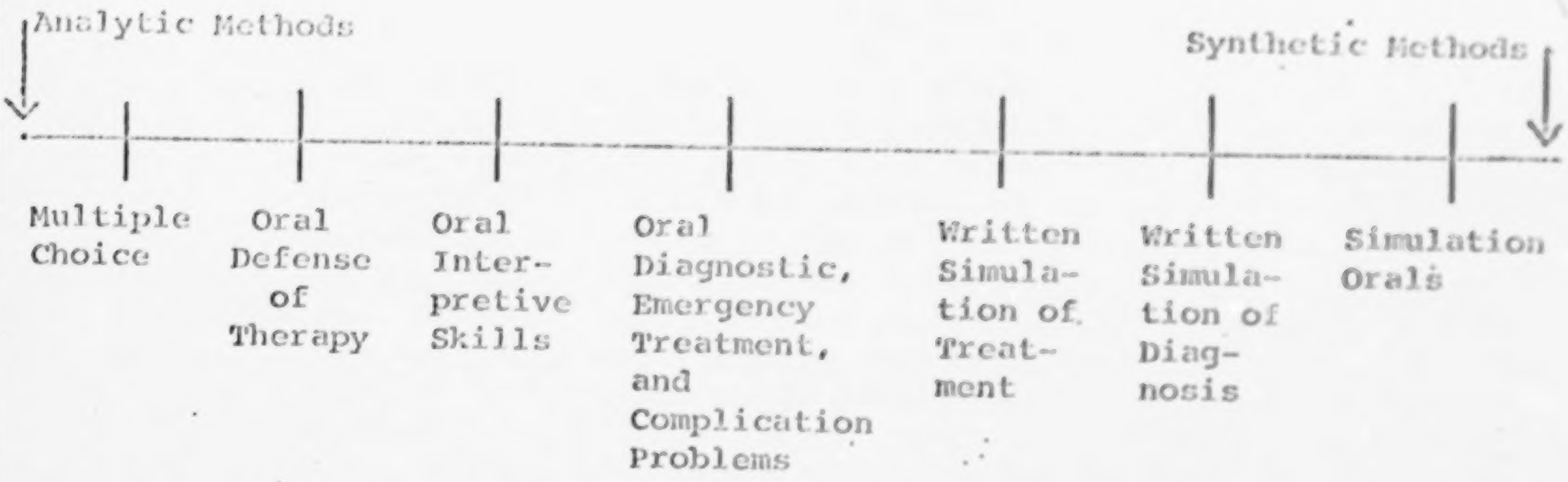
A MODEL FOR THE EVALUATION OF COMPETENCE--- A SYNTHESIS OF THE NEW TECHNIQUES

In the earlier chapters in this section each of the evaluation techniques developed for use in the Orthopedic Training Study was discussed separately in so for as that was possible. However, no evaluation technique can be considered adequately in isolation; in developing a rational system for evaluating complex professional behavior it is necessary to consider the contribution which each possible technique makes to the accuracy and completeness of the overall assessment. This chapter is therefore devoted to an analysis of the interrelationships of the several examination techniques in the prediction of overall competence in Orthopedic Surgery.

Analytic vs. Synthetic Techniques

Any examination represents merely a sample of behavior; as such, it may be designed either to sample some small aspect of behavior that is thought to be an important component of total performance (e.g. muscle coordinations) or, alternatively, to sample wholistically a given slice of total behavior (e.g. the 100-yard dash). The diagram below indicates the location of the several approaches employed in the Orthopedic Study, on a scale ranging from the most analytic to the most synthetic methods of sampling behavior.

As the diagram suggests, the multiple choice technique is the most analytic; it represents an attempt to break total behavior down into its component parts and to sample each bit separately. Such analytic meausres tend to be highly reliable, since with them it is possible to obtain indipendent measures of many small bits of behavior in a relatively short period of time. However, some behavioral traits cannot be validly measured by analytic techniques; for example, reasoning processes can be sampled directly only by such complex techniques as written or oral problem-solving and simulation exercises. Nor can any one case or problem in these more complex methods sample all the qualities of behavior that it may be desirable to measure. For example, in some problems the greatest rewards may go to those who are patient, persistent and moderate; in others, decisiveness in taking radical action may be most valued. The low intercorrelations between scores on different types of oral or of written problems is, in part, attributable to the fact that different problems require different professional


Analytic Methods
Synthetic Methods
Multiple Choice
Oral Defense of Therapy
Oral Inter-pretive Skills
Oral Diagnostic, Emergency Treatment, and Complication Problems
Written Simula-tion of Treat-ment
Written Simula-tion of Diag-nosis
Simulation Orals

qualities; hence the synthetic exercises tend to be more valid, but less reliable. It is for this reason that a variety of examination techniques ranging from analytic to synthetic has been included in the orthopaedic certification examination.

Contribution of Each Evaluation Method to the Prediction of Overall Competence

One might ask, however, if all the modes of examining that were approved for inclusion in the regular certification process are really necessary; i.e., does each make some contribution to the prediction of competence? The multiple regression study of the 1968 Certification Examination is reassuring on this point. The results,* summarized in Table 58A and 58B, indicate that every technique makes some additional contribution to the prediction of Overall Competence as that is defined by supervisors' ratings. Indeed, if the Multiple Choice technique (which has the highest simple correlation with supervisors' ratings of "Overall Competence"), is taken as the starting point, the addition of the score on interpretive skill from the oral examinations and that on proficiency in the written simulations, increases by about 50% the amount of common variance between test scores and supervisors' ratings.

However, it may be argued that even with this increase the value of the combined tests as predictors of overall competence is exceedingly limited since they account for only about 12% of the variance in those ratings. This criticism would have considerable merit if the purpose of the test were to predict the criterion scores. However, such is not the case; their purpose is to identify lack of competence. For this purpose it may be argued that, given the amount of error variance in the ratings (reliability = .31) and the amount of true variance assignable to factors (e.g., surgical skill) none of the exercises was designed to measure, the data strongly suggest that the combined battery of tests predicts about as much of the remaining true variance in overall competence as could reasonably be expected from the limited sample of behavior it is possible to collect in a 6-7 hour examination situation.

Factor Structure of the Test Battery

A second approach to the determination of the value of multiple assessment techniques consists in the analysis of the factor structure of the combined battery of tests and rating scales incorporated in the certification process. In this context, the factor analytic studies made of the 1966 In-Training Examination and the 1968 Certification Examination may be briefly reconsidered here. In the earlier study the three following clearly identifiable factors emerged.**

---

* See also Appendix 15.

** See also Table 6.

TABLE 58 A

RESULTS OF MULTIPLE CORRELATIONAL ANALYSIS
FOR PREDICTING OVERALL COMPETENCE
1968 Certification Examination

N=391

DEPENDENT VARIABLE: SUPERVISOR'S RATING OF OVERALL COMPETENCE R=.365 F=5.28**RELIABILITY=.31

TEST AND SUB TEST SCORES AS INDEPENDENT VARIABLES	PARTIAL r's	F	SIMPLE CORRELATIONS
Interpretation-Interpretation	.119	5.48*	.22
Multiple Choice-Recall	.119	5.46*	.26
Multiple Choice - Problem Solving	.091	3.13	.22
Adult Problem Solving	.077	2.24	.18
Simulations-Attitude	.056	1.21	.15
Trauma Problem Solving	.048	0.86	.18
Written Simulation-Treatment Select Score	.047	0.84	.12
Child Problem Solving	.045	0.77	.14
Written Simulation-Diagnostic, Select Score	.043	0.70	.15
Written Simulation Diagnostic Avoid Score	.023	0.19	.03
Written Simulation Treatment Avoid Score	-.007	0.02	.07

*Significant at .05 level
**Significant at .01 level

TABLE 58 B

INTER-CORRELATIONS AND PERCENTAGE OF COMMON VARIANCE BETWEEN VARIOUS EVALUATIVE TECHNIQUES AND CRITERION DATA

1968 Final Certification Examination Correlations

N=391	Supervisor's Ratings		Test Scores		
Name of Test	Problem Solving	Overall Competence	Multiple-Choice	Oral Score on Interpretive Skill	Written Simulations Proficiency
Correlations:					
Multiple Choice Total	.33	.29	—	.47	.35
Oral Score on Interpretive Skill	.34	.30	.47	—	.26
Written Simulations-Proficiency	.17	.18	.35	.26	—
Weighted Total Score	.39	.35	.80	.81	.36
Percentage of Common Variance:					
Multiple Choice Total	11	8	—	22	12
Oral Score on Interpretive Skill	.12	9	22	—	6
Written Simulations-Proficiency	3	3	12	6	—
Weighted Total Score	15	12	64	64	13

Factor I	A cognitive factor, probably predominantly "recall" behavior, on which the traditional Multiple Choice and oral examinations loaded heavily, and certain scores on the simulated treatment interview and the staff conference had moderate loadings.
Factor II	A reasoning factor, probably predominantly "persistence" in inductive inquiry, on which the written simulations of diagnostic problems loaded heavily and the score on the multiple choice test had moderate loadings.
Factor III	A style or temperament factor, probably involving decisiveness, on which the simulation orals loaded most heavily in a negative direction and written simulations of treatment problems had moderate positive loadings.

The factor analytic study of the 1968 Certification Examination revealed a similar, but substantially more complex factor structure in which 5 factors emerged. One of these additional factors is almost certainly attributable to the inclusion of supervisors' ratings and the second is probably due to the inclusion of certain sub-test scores, the most important of which was the subscore on the written simulations representing skill in avoiding harmful procedures. With the incorporation of these additional measures in the analysis, the following 5 factors were identified:*

Factor I	A general ability factor, on which all ratings have relatively high loadings. It is interesting that among the test scores, the score on Interpretive Skill from the Observation and Interpretation Oral has the highest loading on this factor, suggesting that the confrontations in which chiefs make judgments about residents are often focused around discussion of x-rays and other diagnostic tests or clinical findings.

---

* See Table 7.

Factor II	An information or content factor (similar to Factor I in the 1966 study).
Factor III	A factor related to inductive reasoning, (similar to Factor II in the 1966 study) on which the scores on selection of useful procedures in the written simulations and on the Observation and Interpretive orals have heavy loadings.
Factor IV	A factor of skill in oral communication in which the Simulation Orals have heavy loadings and all orals have at least a moderate loading.
Factor V	A factor related to decisiveness and efficiency (similar to Factor III in the 1966 study) on which scores on avoiding harmful intervention on the written simulation has high positive loadings and that on selecting indicated procedures has high negative loadings.

It is also worth noting that, for the most part, each of the tests included in the 1968 Certification Examination and each of the major scores derived from it had moderate to heavy loadings on several factors. For example, the score on Interpretive Skill derived from the oral examinations showed at least moderate loadings on all 5 of the factors; that on Written Simulations of diagnostic problems on 3 factors, those on the problem-solving orals in adult orthopaedics and in trauma on 2 factors. This factorial complexity of the 1968 examination is especially significant in view of the general tendency for different techniques to emerge as independent factors, partly because each type of test samples some technique-specific abilities (i.e., persons with high verbal facility perform well on oral examinations) and partly because with the inclusion of sub-scores in the analyses (as in the 1968 study) the halo effect (particularly in the supervisors' ratings and oral examination scores) so increases the correlation between related sub-scores that there is a strong tendency for each set of 4 sub-scores to cluster around separate independent factors.

Summary Comment

The data presented in this and preceding chapters strongly suggest that competence in orthopaedics is multifactorial and that a variety of techniques is required to provide valid and comprehensive information on candidates applying for certification. These data also highlight the necessity of developing a system for scoring and reporting test results that takes full cognizance of the philosophical, psychological and psychometric issues discussed above and which is, at the same time, practical, feasible and acceptable to both the candidates and the Board. The method which was adopted--a profile system--is described in the following chapter.

CHAPTER X:

THE APPLICATION OF THE PROFILE TECHNIQUE TO THE PROBLEMS OF CERTIFICATION

Criteria for Designing the System

A profile technique of summarizing and reporting the results of the examination system described in the preceding sections was developed. This system was designed to meet three criteria which are often violated in traditional methods of utilizing test data for purposes of certification. These criteria can be summarized as follows:

(1) Competence in orthopaedic surgery is multifactorial in nature; it therefore follows, that strength in one area cannot be allowed to compensate completely for weakness in another.

(2) The unit in terms of which competence is assessed should be based on performance factors, not individual tests; only when each test technique measures a different trait, should scores on individual tests be considered separately.

(3) Ideally, the level of satisfactory performance on a certification examination should be determined prior to its administration and should be based on absolute standards not on relative standing in the distribution of candidate scores.

Procedures in Implementing the System

In order to meet these criteria the following four performance factors* were identified as the units in terms of which certification decisions were to be made on the 1968 Orthopaedic Certification Examination: Recall of factual data, Analysis and interpretation of clinical data, Problem-solving ability and Attitudes toward patients and colleagues.

---

* See Table 31

Sub-scores on each factor were derived from each of the oral and written tests included in the 1968 test battery. These sub-scores were converted to a 12-point scale with 3.5 defined as the minimum passing level. Sub-scores, weighted as shown in Table 59, were combined to obtain the four factor scores and the total score, from which a profile similar to that shown in Table 60 was derived for each candidate.

Prior to the administration of the examination, the following tentative guidelines for determining certification were adopted:

1. All candidates scoring below 3.5 (the Failing area) on Problem Solving, Interpretation, or Recall should NOT be certified.

2. All candidates scoring between 3.5 and 6.4 (the Marginal area) on the TOTAL should be reviewed and ground rules established for the disposition of each case.

3. All candidates scoring above 6.4 (Good or Excellent) on the TOTAL and above 3.4 on every factor should be certified.

In adopting these guidelines for use with a profile system of scoring the Board formally recognized the multifactorial nature of orthopaedic competence, since the four performance factors rep sented a distillation of the 94 components of competence derived in the critical incident study and since the Board continued to require (as previously) that each candidate submit data attesting to his surgical skill and professional ethics as a condition for admission to the certifying examination. (Ratings on The Candidate Evaluation Forms required for each applicant represented an attempt to systematize the collection of data on the latter qualities.)

Second, the derivation of scores on each factor from a variety of techniques assured the maximum reliability of each factor score. While a score on an individual examination may be so unreliable as to be virtually meaningless, a low score derived across a number of techniques offers reasonable certainty that the individual is inadequate on that performance factor. Third, by relying on a pre-established "Minimum Passing Level" and pre-determined ground rules the Board assured that "pass-fail" decisions would be based on absolute standards rather than arbitrary decisions.

TABLE 59

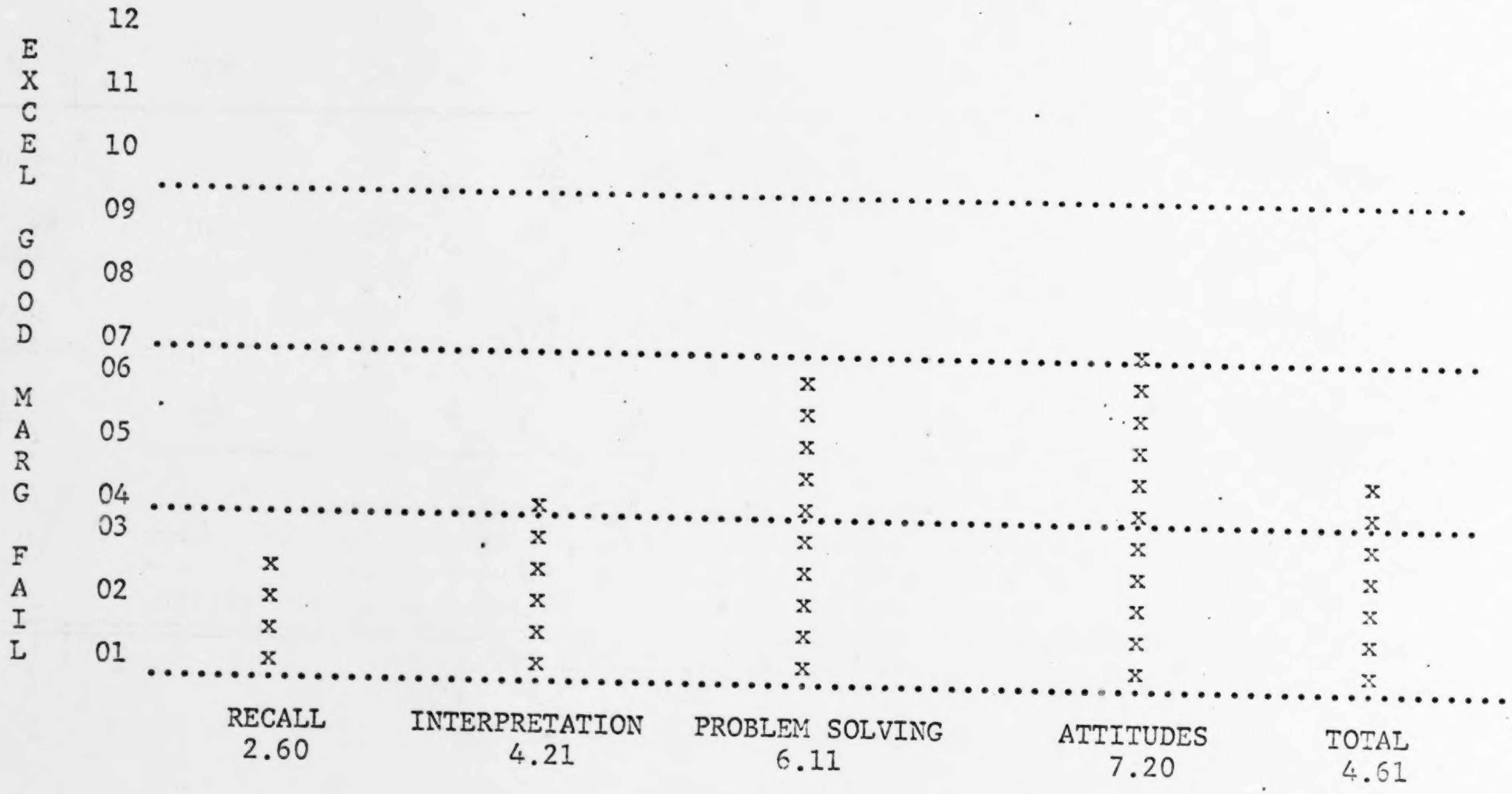
WEIGHTS ASSIGNED EACH SUB-SCORE
1968
FINAL CERTIFICATION EXAMINATION

	Multiple Choice Exam	Written Simulation Exercises	Adult Problems Oral	Child Problems Oral	Trauma Problems Oral	Observation and Interpretation Oral	Simulations Orals Examiner A	Simulations Orals Examiner B	TOTAL
Recall	.28	.02	.01	.01	.01	.01	.005	.005	.35
Interpretation	x	.01	.035	.035	.035	.075	.005	.005	.20
Problem Solving	.06	.10	.06	.06	.06	.01	.005	.005	.35
Attitudes	x	.01	.005	.005	.005	.005	.035	.035	.10
Total	.33	.14	.11	.11	.11	.10	.05	.05	1.00

TABLE 60

THE AMERICAN BOARD OF ORTHOPAEDIC SURGERY
1968 CERTIFICATION EXAMINATION

CANDIDATE PROFILE NUMBER X



Results

In January, 1968 the first certification examination to which this profile scoring system would be applied was administered to 838 candidates of whom 54 were retaking only a segment of the examination. The nature of the remaining candidate population is reported in Table 61 below:

TABLE 61

CANDIDATE POPULATION

	Graduates of U. S. Medical Schools	Graduates of foreign medical schools	Total
Admitted to final certifying exam without prior examination (due to change in rules of eligibility)	54	22	76
Previous examination experience but no prior examination failure	503	27	530
Previous examination failures	128	50	178

Scores and sub-scores on each test and each factor (Recall, Interpretation, Problem-Solving and Attitudes) were computed and profiles drawn for each individual. Univariate statistics were derived for the total population and for the various sub-groups described above. These results are reported in Tables 62 and 63. The interrelations among scores and sub-scores and between them and the multifactorial rating of candidates by their chiefs was analyzed by correlational and multiple regression techniques. Tables 64 and 65 below present the results of that investigation.

Table 62, below, reports the results of performance on the total examination and on each factor for the over 500 graduates of U. S. medical schools who were taking the certification examination for the fir time.

TABLE 62

DISTRIBUTION OF SCORES ON 1968 FINAL CERTIFICATION EXAMINATION FOR U.S. GRADUATES WHO WERE TAKING THE FINAL CERTIFICATION EXAMINATION FOR THE FIRST TIME

	Scores	Recall		Interpretation		Problem-Solving		Attitudes		TOTAL	
		N		N		N		N		N	
Excellent	11.5	1	100.0					7	100.0		
	11.0	0						16	98.8		
	10.5	3	99.8	4	100.0			56	96.0		
	10.0	5	99.3	9	99.3	1	100.0	49	86.3		
	9.5	6	98.4	29	97.7	8	99.8	54	77.7	6	100.0
Good	9.0	9	97.3	61	92.7	35	98.4	71	68.3	5	99.0
	8.5	18	95.8	70	82.1	44	92.3	64	56.0	27	[illegible]8.1
	8.0	6	92.7	80	69.9	91	84.7	58	44.9	46	93.4
	7.5	29	91.7	79	56.0	105	68.9	52	34.8	84	85.4
	7.0	35	86.6	63	42.3	111	50.6	47	25.7	98	70.8
	6.5	29	80.5	68	31.3	70	31.3	34	17.6	93	53.7
Marginal	6.0	39	75.5	53	19.5	52	19.1	16	11.7	103	37.6
	5.5	45	68.7	20	10.3	31	10.1	17	8.9	55	19.7
	5.0	48	60.9	13	6.8	16	4.7	14	5.9	35	10.1
	4.5	68	52.5	15	4.5	8	1.9	17	3.5	15	5.0
	4.0	125	40.7	8	1.9	3	0.5	3	0.5	6	1.4
	3.5	67	20.0	1	0.5					2	0.3
Failing	3.0	32	7.3	2	0.3						
	2.5	5	1.7		0						
	2.0	3	0.9		0						
	1.5	1	0.3		0						
	1.0	1	0.2		0						
Mean		5.15		7.42		7.18		8.31		6.63	

After reviewing the data on the distribution of scores shown in Table 62 the tentative guidelines listed above were accepted with the following minor modifications:

1. Since the Interpretive score was based almost exclusively on 4 of the 5 orals, and primarily on one of them, it was decided that a failing score on this factor alone would not be sufficient cause to withhold certification. This change affected only 2 candidates.

2. It was decided to certify everyone whose Total score was clearly satisfactory, i.e., 6.5 or above. This decision was made on the ground that the Total score was significantly more reliable than scores on the independent performance factors. This decision resulted in the certification of one candidate who otherwise might have failed.

In addition the following guidelines were developed for dealing with the 38% "marginal" candidates in the normative group:

1. Certification would be withheld from any candidate whose scores were "Marginal" on the Total AND on any 3 of the 4 performance factors.

2. Certification would be withheld from any candidate whose scores were marginal on BOTH Recall and Problem-Solving.

In arriving at these decisions, the American Board of Orthopaedic Surgery took cognizance of the fact that 76% of the candidates scored at marginal or failing levels on the Recall factor and 20% scored at these levels on Problem-Solving. While these results may raise some doubt about the appropriateness of the pre-determined standards on Recall, they are compatible with the view that the store of information readily accessible to a large number of candidates is marginal for optimal orthopaedic practice, and that in such cases certification should be awarded only to those whose problem-solving and other skills are "Good" or "Excellent." In short, the Board adhered closely to the guidelines promulgated BEFORE the examination and, for the first time, implemented a system based on absolute, rather than relative standards. The effects of the adoption of these guidelines on the various subgroups within the candidate population are shown in Table 63 below.

TABLE 63

PER CENT FAILURE RATE AMONG DIFFERENT CANDIDATE POPULATIONS

	Graduates of U. S. Medical Schools	Foreign Graduates	Total
No previous examination experience	24	59	34
No previous examination failures	18	48	19
Previous examination failures	59	80	65
Total	26	67	31

In a system such as this in which decisions are made on the basis of pooled judgments of numerous observers, each sampling (as objectively as possible) a "bit" of candidate behavior it is important to consider the interrelations among examination scores and between them and the ratings made by senior staff who are familiar with the candidate. Table 64 summarizes such data.

TABLE 64

INTERCORRELATIONS OF EXAMINATION SCORES AND RATING FACTORS

(N=391)

	Examination Scores				
	Recall	Interpretation	Problem Solving	Attitudes	Total
Examination Scores					
Recall	x				
Interpretation	.42	x			
Problem-Solving	.48	.69	x		
Attitudes	.22	.46	.40	x	
Total	.84	.77	.83	.51	x
Rating Factors:					
Problem-Solving	.32	.36	.30	.21	.39
Patient Relations	.13	.22	.17	.17	.21
Overall Competence	.29	.38	.29	.17	.35

As might be expected, these data indicate that within the examination, scores on "Interpretation" and "Problem Solving" are most closely related to each other, and those on "Recall" and "Attitudes" are least so. Secondly, though the correlation between the examination scores and training chief's ratings are generally low (due in part to unreliability, especially of the latter) the data reveal a significantly higher correlation between the chief's rating of candidate's problem-solving skills and the examiner's rating of the cognitive components of competence, than between the chief's and the examiner's assessment of the affective components of competence.

TABLE 65

RESULTS OF MULTIPLE CORRELATIONAL ANALYSIS USING EXAMINATION PERFORMANCE FACTORS AS INDEPENDENT VARIABLES

1968 Final Certification Examination

N=391

Dependent Variables (Rating Factors)	Correlation with Total Examination Scores	Multiple R with all 4 Performance Factors	F	Independent Variables (Examination Scores)	Partial r	F	Simple r
Information Gathering	.35	.36	14.32**	Recall	.15	8.67**	.28
				Interpretive Skill	.15	8.31**	.32
Problem Solving	.39	.40	18.68**	Recall	.18	12.82**	.32
				Interpretive Skill	.17	11.23**	.36
Clinical Judgment	.32	.34	12.36**	Interpretive Skill	.14	7.79**	.30
				Recall	.11	4.82*	.24
Surgical Technique	.23	.26	8.04**	Interpretive Skill	.15	8.64*	.25
Patient Relations	.21	.24	5.83**	Interpretive Skill	.11	5.11*	.22
				Attitudes	.08	5.21	.17
Continuing Responsibility	.26	.27	6.24**	Interpretive Skill	.13	6.24*	.26
Emergency Care	.26	.28	8.22**	Interpretive Skill	.15	9.36**	.27
Colleague Relations	.22	.23	5.52**	Interpretive Skill	.10	3.58	.21
				Recall	.07	2.08	.17
Ethics	.24	.25	6.32**	Problem Solving	.09	3.47	.23
				Recall	.08	2.55	.18
Overall Competence	.35	.36	14.40**	Recall	.16	9.75**	.29
				Interpretive Skill	.13	7.68**	.31

NOTE: Data on factors with F ratios below 2.00 are not reported
* Significant at .05 level
** Significant at .01 level

Finally, it should be noted that the score on Interpretation is more highly correlated than any other examination factor with the chief's rating of Overall Competence; a result that probably reflects the fact, noted earlier, that many of the staff-resident encounters are concerned with the interpretation of X-Ray and other diagnostic studies.

Finally, the results of the multiple correlational analysis using examination performance factors as independent variables (see Table 65) provide further evidence as to the validity of the profile technique of scoring and reporting. The data indicate that a single score by itself, is not sufficient to describe competence. Second, they suggest that the components of competence measured by the examination scores on "Recall" and "Interpretation" are about equally decisive, and that both are more important than the components of competence, measured by the "Problem-Solving" and "Attitude" scores, in contributing to the chief's judgment of overall competence and of many of the cognitive skills which he rated. This finding is, in itself, significant in considering the basis on which training chiefs evaluate resident competence. Third, the aspect of competence measured by the "Attitude" score contributes significantly only to the prediction of the chief's rating of "Effectiveness in Patient Relations." Finally, the empirical weightings of the various scores (multiple regression equation) are, for the most part, similar to the pre-assigned weights (Table 59) which had been developed on purely logical criteria for purposes of computing the factor scores and the composite total score on the 1968 Certifying Examination.

Summary Comment

The profile system of scoring and reporting test data, as developed by the American Board of Orthopedic Surgery for use in its 1968 examination program, constituted a major innovation in procedures for specialty certification. Among the most important characteristics of this system are (1) the provisions developed for obtaining appropriately weighted pooled judgments derived from a number of sources about each major performance factor that contributes to competence, (2) the methods developed for determining and applying absolute standards in judging performance, (3) the nature of the feedback the system provides to candidates, the training chiefs and the Board and (4) the self-corrective mechanisms which this feedback system stimulates.

SECTION THREE

CURRENT STATUS
AND
PROJECTED NEXT STEPS

CHAPTER XI:

OUTCOMES OF THE ORTHOPAEDIC TRAINING STUDY

To provide a context in terms of which the major outcomes of the 4-year Orthopedic Training Study can be summarized and evaluated it would be well to recall briefly both the immediate objectives and the long-range goals for which that research project was designed. Since the study was initiated as a direct consequence the continuing concern of the American Board of Orthopedic Surgery for systematic improvement of its certification procedures, the development of more valid and more reliable techniques of assessing professional competence in orthopedics constituted the immediate research aim. This, in turn, entailed the development both of a methodology and of specific instruments for the three-fold purpose (a) of defining professional competence in operational terms, (b) of analyzing existing certification techniques and (c) of constructing and validating more appropriate ones. Second it was clear from the outset of the study that the problems faced by the American Board of Orthopedic Surgery are not unique to it; they are common to all groups responsible for setting professional standards and are especially urgent in all of the health fields where rapid scientific advances combined with increased popular demand for more and better professional services have excerbated the problems of setting and maintaining standards. Thus, the development of a model for professional self study became an intermediate goal of the study. Finally, evaluation was viewed by both the research team and the Board as an integral part of the educational process and thus, as a prerequisite for increasing the efficiency and enhancing the effectiveness of professional training, these being the long-term objectives of a project designed to contribute ultimately to the better utilization of scarce manpower resources.

Immediate Outcomes

Materials

The specific instruments developed in this project, together with the findings regarding each, have been described in detail in previous chapters; here they may be briefly reviewed under the following heading: rating forms, tests and test manuals, forms for profile scoring and reporting of test results and observational forms for process analysis of tests.

In summary, two types of rating forms were developed: one for use in scoring oral or practical examinations; the other, for recording assessments of habitual behavior, the latter developed to obtain evidence on those skills and attitudes that cannot be adequately sampled in limited "test" situations, as well as to obtain data for studies of the concurrent validity of other instruments. Irrespective of the purpose for which they were devised, the setting in which they were used or the group to whom they were applied, all rating forms retained for incorporation in the regular certification procedure, shared three important characteristics: (a) they specified distinct aspects (factors) of performance which the observer was asked to rate; (b) they untilized a 12-point scale on which points were grouped into four levels of performance with each level (usually) defined in terms of absolute, rather than relative, standards of performance; and (c) either on the rating scale itself, or in an accompanying manual of instructions, each factor, was operationally defined and the behavior representing each end of the scale was described and illustrated concretely. These three characteristics of the rating forms are regarded as of primary importance.

The specific test materials developed for use in this study included from 1 to 5 forms of standardized written and oral exercises of a variety of types incorporated on the in-training and certification examinations scheduled during the research period. However, unlike most other research of the same nature, the actual instruments are probably of less significance than the new techniques and associated manuals devised during the course of the study. Among these, the ones of greatest long-run value appear to be the following: Written simulations of both diagnostic and theraputic problems in patient management requiring sequential analysis and decision, oral simulations of physician-patient and physician-colleague encounters, oral exercises sampling interpretive and problem-solving skills, the latter as applied to diagnostic and treatment problems in both emergency and chronic disease situations. All of these new techniques share 3 common characteristics: first, each was devised to sample a clearly defined segment of competence identified in the critical incident study as one of the requisites of effective professional performance; second, whether in oral or written format all exercises are based on standardized case materials presented in a manner that elicits as directly as possible the behavior each was designed to sample; third, methods of recording (and/or observing) the examinee's responses to the test situation are such as to afford reasonable objectivity in scoring and to facilitate the application of predetermined and explicitly defined standards in judging the candidates' performance. Finally, for each of the new test techniques a brief manual has been prepared outlining procedures for developing appropriate test materials, for administering and scoring that type of exercise, and for setting minimal acceptable standards and interpreting performance on it.

As with the test materials, the specific forms and procedures developed for profile scoring and reporting of individual candidate performance in orthopedic surgery, while available, are probably of less generalized significance than the overall rationale and methodology that was developed as a basis for deriving the specific forms. The primary advantages of the system as such appear to be threefold: (1) by combining numerous "bits" of information obtained from samples of candidate performance in a variety of types of settings the composite result is maximally reliable; (2) since each point of the profile is based on behavioral factors (e.g. interpretive skill) rather than test techniques (e.g. score on multiple choice questions) the "picture" of competence yielded by the profile corresponds more closely to operational definitions of competence employed by training chiefs and colleagues in evaluating physician performance than does the picture which emerges from more conventional scoring techniques; (3) finally, the nature of the profile provided on each candidate greatly facilitates the application of appropriate standards despite the great variety in patterns of professional competence characteristic of any population of applicants.

In addition to the specific instruments developed for use in assessing individuals, the two observational forms developed for describing and evaluating the tests themselves deserve special mention. The somewhat different forms utilized in the studies of the traditional and the new type oral examinations shared two important characteristics: (1) insofar as possible the observer was required to identify the nature of candidate (and/or examiner) behavior elicited during the test period; he was not asked to evaluate the "quality" of the examination; (2) the observer was instructed to make such a descriptive recording for each unit of behavior rather than to furnish an account based on overall impressions. This technique of classifying and recording bits of information appeared to maximize the objectivity and the reliability of the observations while furnishing the basic data essential for a subsequent qualitative judgment regarding the content validity of each type of oral.

Findings

No attempt will be made here to provide a comprehensive summary of the specific results discussed fully in earlier chapters; at this point it would seem most appropriate merely to outline briefly the major trends observed with respect to each of the following general categories of findings: (1) those concerned with the reliability and validity of each of the new test techniques; (2) those concerned with variations in patterns of performance associated with age, level of education and experience, nature of practice setting and the like (obtained from cross-sectional studies of different population samples); and (3) those concerned with changes in professional achievement associated with increased education and /or experience (as obtained from longitudinal studies of the same population sample).

Studies of the reliability of the various measures indicate that reasonably lengthy multiple choice examinations and written simulations of patient management problems are highly reliable measures when reliability is defined as degree of internal consistency; however, there is sufficient variation in approach to different types of patient management problems to indicate that a number of problems, both diagnostic and therapeutic, ranging from emergency to comprehensive care situations, and sampling a number of clinical entities should be included in one examination in order to generalize the results to a universe of varied clinical problems. As regards the oral examinations, all achieve a level of interrater reliability sufficient to justify their inclusion in a battery of tests, provided single scores from individual orals are not used independently to determine passing and failing. Further the traditional orals. as well as the new simulation and interpretive orals reach acceptable levels of sampling reliability provided scores are not treated independently. However, sampling reliability of the problem-solving orals is sufficiently lower to suggest the necessity of using several cases even when this examination is employed as part of a test battery.

Findings with respect to validity of the various techniques are somewhat more difficult to summarize: Content validity, studied by both process and observational analysis, was judged to be significantly higher for all of the newer techniques than for the more conventional techniques. Construct validity was studied by exploring the congruence between hypotheses about the performance of groups at various levels of training and experience, and their actual performance on a given test. As might be expected, these studies indicate that despite essentially complete overlap in the range of scores for groups at different levels of training, mean scores on most tests differ in the expected direction. Amount of training is most highly correlated with performance on tests that measure general orthopaedic information or decisiveness about therapy and least correlated with scores on tests designed to assess thoroughness of diagnostic work-up or the ability to relate to patients. These results are consistent with other information about the general nature and relative emphasis that characterize most training programs. Further, studies of the influence of experience and type of practice on responses to the written simulations of patient management problems reveal the same types of relationships as described in observational studies of practitioner performance. Concurrent validity of the various measures was investigated through correlational and factor analytic studies of the interrelation among scores on different types of examinations and between them and supervisor's ratings of performance. These studies revealed that there was considerable overlap in the conventional written and oral examinations and that the newer techniques appeared to measure aspects of competence not previously sampled. The intercorrelations of scores on the new techniques (when corrected for attenuation) yielded a factor structure compatible with hypotheses as to the interrelationships among aspects of competence each was designed to measure. However, correlations between test scores and supervisor's ratings were often disappointly low, though the patterns of these correlations matrices were as predicted when due allowance was made for the sometimes exceedingly low reliabilities and often very large errors due to "halo" effects that characterized the ratings. The predictive validity of the newer certification techniques is to be investigated in the proposed ten year follow-up study outlined below.

Among the findings in regard to variations in performance associated with age, level of training, practice setting and like, the following are of greatest significence: the relatively slight differences found between groups at different levels of training on several of the achievement measures, the consistent tendency for performance of practitioners on the written simulations to decline with age and with remotenesss of affiliation with a teaching institution, and the striking diminution in diagnostic thoroughness associated with increased amounts of training and experience.

Finally, data are now available from repeated administrations at one year intervals of parallel types of achievement tests to a population on which substantial amounts of biographical information are also available. These data are now being analyzed as a preliminary step in a newly initiated study of the relationship between patterns of growth in professional competence and specified training variables. Preliminary analysis suggests that in the absence of special factors, increased individual achievement associated with increased amounts of training will be most readily demonstrable in terms of the amount of specialized information the individual can recall, the level of surgical skill he displays and the decisevness with which he embarks on a plan of treatment, and that least achievment will be demonstrable in the areas of professional habits and attitudes. The extent to which these general trends are modified by variations in the nature of the training program is the subject of the forthcoming study.

Intermediate Outcomes

In planning the Orthopedic Training Study it was hoped that experience with certain of the methodologies developed during the course of the investigation could provide a basis for developing a generalized model for professional self-study applicable to other groups in the health professions. In this context attention should therefore be redirected both to the rationale of the study and to the more important characteristics of the organizational structure evolved for implementing it.

The approach underlying the study has been explictly treated at numerous points throughout this report and particular methodological developments of general interest are noted above among the materials immediately available from the study. This discussion is therefore limited to identification of the following major features of the study rationale that appear to be of greatest general applicability: (1) The use of an empirically derived, behavioral definition of the essential components of professional competence to guide every stage of the study; (2) the focus on the nature of the behavior to be sampled in the design, scoring, and evaluation of new assessment techniques, (3) the provision for systematic feedback of results to members of the profession responsible for maintaining standards and for making decisions regarding training and certification policies to implement these standards; (4) the utilization of reliable and valid assessment not as an end in itself nor even as a means of maintaining professional standards, but as a prerequisite to sound educational experimentation and as an indispensable part of any educational program.

Any intensive professional self-study necessarily entails an interdisciplinary approach; however, though they have led to recommendations for modification of educational programs, many such interdisciplinary studies have failed to produce any solid long-term accomplishment. It is for this reason if for no other that in developing a generalized model for professional self-study it is of considerable importance to consider the nature of the organizational structure required both to promote effective utilization of the various types of expertise needed during the course of the investigation and to facilitate the ultimate transfer of responsibility for implementing the findings of the study, from a specially appointed research team to the regularly constituted policy-making bodies of the profession. Experience suggests that the following pre-existing conditions were especially favorable in the present study:

(1) The quality of leadership in the orthopedic profession vis-a vis educational issues. Long before the initiation of the study the leadership in the orthopedic profession had evidenced continued concern about problems of education and evaluation; for example, it was the first medical specialty to introduce an in-training examination to assist in monitoring resident progress; it had previously sought guidance from educational specialists and it was a request from the Board for a review of its certification procedures that stimulated the conversations eventuating in the research proposal represented by this study.

(2) The history of involving a large number of orthopedists in the regular certification procedures of the Board. For a number of years the Board had made it a regular practice to utilize the services of over 200 senior members of the specialty (including virtually all of those with major responsibilities for training programs) in developing written examination materials and adm[illegible]istering oral examinations; during this period the Board had developed regular procedures for recruiting and orienting new

examiners, for consulting with both them and candidates about the conduct of the examinations and methods of improving them, and for feeding back to this cadre information about the results of the examination.

(3) Previous experience of the Center in interdisciplinary research in medical education. Prior to the present study all members of the Center staff had had long experience in interdesciplinary research in education and many of them specifically in medical education; the basic research staff itself included specialists from both medicine and education. Both of these circumstances greatly mitigated communication difficulties often encountered in interdisciplinary research.

(4) The methods evolved for conducting this study. From its exception the study was viewed as a joint undertaking of the Board and the Center; the research proposal was cooperatively developed; provision was made from the outset for periodic joint review and planning, for allocation of specificed responsibilities between the Board and the Center, for budgeting to include orthopedists recruited by by the Board both as a part of the regular research staff and as full-time consultants on special aspects of the study for periods ranging from a few days to several weeks, for travel and meeting funds to support the efforts of numerous Task Forces appointed by the Board to work with the research staff on specific problems, for regular communications on the nature and progress of the study to the profession at large either directly from the Board or via it, from the research staff and, finally, for the Board to take over the implementation of new policies and procedures once they had been developed to an operational level by the research staff.*

That this transfer of responsibility has occurred is evidenced by the fact that either directly or indirectly the study has influenced introduction of, or planning for, the following modifications in Board certifying policies or procedures:

(1) In accord with the components of competence defined in the critical incident study the Board has established an examination blueprint specifying the cognitive skill, attitudinal processes and the subject matter content to be evaluated, and the weight to be assigned each in the certifying examination; this blueprint has been adopted for use in defining the specifications for all examinations under the Board's juristiction.

2. The Examination Committee of the Board has established regular procedures for maintaining and updating the classification of materials in the examination pool in accord with the categories of the blueprint.

---

* See Appendix 29 for a detailed listing of joint activities of the Center and Board of a special nature.

(3) The Board has established regularly constituted task forces charged with responsibility for developing, reviewing, refining and upgrading materials for the multiple choice, written simulation and oral components of the examination.

(4) The Board has developed a greatly improved system of preparing and reviewing written examinations for certification to assure conformity to specifications of the blueprint.

(5) The total certifying process has been redesigned to yield evidence on the following aspects of professional competence:

I Surgical skill
II Professional habits and attitudes
III Ability to recall information
IV Ability to interpret and analyze data
V Ability to solve problems (including clinical judgment)
VI Ability to relate effectively to patients and colleagues

Evidence on Factors I and II is to be gathered primarily by use of questionnaires and rating scales developed during the present study. The Multiple Choice Examination has been redesigned to yield evidence on Factors IV and V, as well as Factor III. Written simulations (Patient Management Problems) are to be utilized as a regular part of the certifying examination to yield evidence on Factor V (and where relevant, on Factor IV). The Oral Examination has been re-designed to include three half-hour examinations designed to assess interpretive skill and one half-hour examination designed to assess skill in relating to patients (simulated physician-patient encounters). Each component of the oral examination is administered by trained examiners utilizing previously prepared standardized case materials and is scored on standard, objective rating forms.

(6) The previous scoring system, in which each examination was treated independently, has been replaced by a profile of performance in which evidence from several sources is combined to yield the most reliable assessment of each factor.

(7) A program for training examiners in the development of standardized materials and in the administration and scoring of oral examinations has been instituted.

(8) Provision has been made for the establishment and up-dating of a data bank to contain all "bits" of information which the certification process yields.

9. The Board has established its own office of Education and Evaluatio staffed with a full-time director to implement these revised policies and procedures.

Long-Term Outcomes

It is, perhaps fortunately, too early to try to assess the extent to which the study achieved its long-range objectives of contributing to increased effectiveness and efficiency in professional training and by that means to the better utilization of score man-power. At this stage, however, three specific outcomes with long range implications should be noted: The first, and in the long-run perhaps one of the most significant, is the increased sensitivity to and sophistication about educational principles and problems achieved by the large number of Orthopedists associated in one way or another with the study, as evidenced by innumerable bits of annecdotal data about individual changes in educational philosophy and/or practice. Second, as one direct consequence of the study the profession has been furnished with a technology, a methodology and, to a certain extent, a pool of trained personnel which facilitate the continued self-study that leaders in the speciality appear to be highly motivated in pursuing. Finally, and most concretely, the Board by its decision to eliminate time and distribution requirements in specified training programs for a stated period, has cleared the way for initiating second extended joint study devoted directly to educational experimentation. This study, as described more fully in the next and final chapter, could not have been undertaken in the absence of a reasonably comprehensive and valid evaluation system for purposes of individual certification and program assessment, such as that which has now been made operational.

CHAPTER XII:

A LOOK TO THE FUTURE

The summary of current status suggests the logical next steps to be taken in the immediate future. In addition to the provisions for continued improvement in the regular certification procedures of the Board and the further analysis of currently available longitudinal data on the relationships between aptitude and achievement and between achievement patterns and training variables discussed previously, two specific extensions of the present research are clearly indicated. Specifically, these are the design of a follow-up study to assess the predictive validity of the newly established certification procedures and the initiation of systematic experimental investigation of methods of improving professional training.

A Proposed Ten Year Follow-Up Study*

In the initial plan of the Orthopedic Training Study it was stipulated that a ten-year follow-up would be made of candidates applying for certification during the period of the research to determine the differences, if any, in the quality of health care delivered by successful and unsuccessful candidates. The detailed proposal for such a study is presented in Appendix 30 . Briefly it is recommended that samples of successful and unsuccessful candidates be drawn from the populations applying for certification in the three years just prior to the initiation of the study and in the three years immediately following full implementation of the new certifying procedures. It is suggested that on each of these samples specified performance data be collected by means of self-report questionnaires, patient logs, confidential assessments from chiefs of staff, peer ratings, review of hospital charts and direct observation in office or hospital settings, and that these data be analyzed to determine the relationship between current performance and that on various certifying instruments. Such analysis should yield valuable information not only on the predictive validity of the certifying process, but also on the changes over time in patterns of competence of significance in improving current residency training.

*For a detailed outline of this study see Appendix 30 .

Educational Improvement Studies

The related interests of the American Board of Orthopaedic Surgery, the Musculo-Skeletal Committee (NRC-NAS) and the Center for the Study of Medical Education are joined in an experimental study of educational innovation building on and extending the current investigation in order to achieve the following broad objectives:

1. To provide a model of individualized graduate education in medicine in which the demonstration of individual competence, rather than the fulfillment of rigid time and content requirements, marks the end point of formal training.

2. To document the nature and variations of orthopedic training in the United States.

3. To devise and test methods for increasing the efficiency and effectiveness of orthopedic training.

4. To determine the relationships between input, training and output variables.

5. To develop mechanisms that will facilitate continuing institutional self-study of training programs.

6. To develop a pool of educational specialists in orthopedics who can provide continuing leadership in the field.

In the joint study initiated in July, 1968, it is proposed to accomplish these objectives in a three-stage study devoted first to an intensive investigation of the nature of current training experience, second to controlled experimental modification in educational acitivities in selected training programs, and finally to an analysis of the interrelations among input, output and training variables.

In the first stage of the study data about each program in both the control and experimental groups will be collected with regard to the following:

1. Program organization--including schedule of resident rotation, the personnel who supervise training, the facilities and resources to support the training.

2. Program objectives--the mechanism of their establishment, review and communication to staff and residents.

3. Program operation--activities and responsibilities of a resident sample, the nature of instructional procedures both formal and informal, the nature of feedback to residents of their individual strengths and weaknesses as training progresses.

4. Program evaluation--the mechanisms employed to accumulate data about resident progress and program effectiveness, and the utilization of these data in continuing program review.

5. Program perceptions--identification of similarities and differences among residents and staff in the perception of purposes, procedures and effectiveness.

It is anticipated that such intensive review will quickly identify areas in which new organization of training systems, or utilization of alternative instructional modes would predictably increase either efficiency or effectiveness of training. In the experimental institutions this would lead in the second stage of the study to introduction of specific instructional innovations as well as to more fundamental modifications in program organization and in those more subtle and pervasive factors of staff-trainee interactions that influence the basic climate of learning.

As changes both in specific methodology and in the general climate for learning are introduced their effect upon resident achievement will be assessed through both cross sectional and longitudinal studies to which the third stage of the study will be increasingly devoted.

The research outlined above is a direct outgrowth of the current study. It is with the view of the future provided by the initiation of the new study in July, 1968, that the report of the first Orthopaedic Training Study is most fittingly terminated.

APPENDIX

Appendix 1

MATERIALS FOR THE EVALUATION OF PERFORMANCE
in
MEDICINE

DETERMINATION OF OBJECTIVES

prepared by

The Evaluation Unit
Center for the Study of Medical Education
University of Illinois, College of Medicine
January, 1967

ACKNOWLEDGEMENTS

The following definition of the critical performance requirements of Orthopaedic Surgeons was derived from a joint study conducted by the American Board of Orthopaedic Surgery and the Center for the Study of Medical Education, University of Illinois College of Medicine and is reprinted by their special permission.

The Critical Incident Study was carried out with the assistance of The American Institutes for Research, Pittsburgh, Pennsylvania.

163

ORTHOPAEDIC TRAINING STUDY
AMERICAN BOARD OF ORTHOPAEDIC SURGERY

AND

CENTER FOR THE STUDY OF MEDICAL EDUCATION
UNIVERSITY OF ILLINOIS

Critical Performance Requirements for Orthopaedic Surgeons
(derived from The 1964 Critical Incident Study)

I. Skill in Gathering Clinical Information
  A. Eliciting Historical Information
    1. Obtaining adequate information from the patient
    2. Consulting other physicians
    3. Checking other sources
  B. Obtaining Information by Physical Examination
    1. Performing thorough general examination
    2. Performing relevant orthopedic checks

II. Effectiveness in Using Special Diagnostic Methods
  A. Obtaining and Interpreting X-rays
    1. Directing or ordering appropriate films
    2. Obtaining unusual, additional or repeated films
    3. Rendering complete and accurate interpretation
  B. Obtaining Additional Information by Other Means
    1. Obtaining biopsy specimen
    2. Obtaining other laboratory data

III. Competence in Developing a Diagnosis
  A. Approaching Diagnosis Objectively
    1. Double-checking stated or referral diagnosis
    2. Persisting to establish definitive diagnosis
    3. Avoiding prejudicial analysis
  B. Recognizing Condition
    1. Recognizing primary disorder
    2. Recognizing underlying or associated problem

IV. Judgment in Deciding on Appropriate Care

A. Adapting Treatment to the Individual Case

1. Initiating suitable treatment for condition
2. Treating with regard to special needs
3. Treating with regard to age and general health
4. Attending to contraindications
5. Applying adequate regimen for multiple disorders
6. Inventing, adopting, applying new techniques

B. Determining Extent and Immediacy of Therapy Needs

1. Choosing wisely between simple and radical approach
2. Delaying therapy until diagnosis better established
3. Testing milder treatment first
4. Undertaking immediate treatment

C. Obtaining Consultation on Proposed Treatment

1. Asking for opinions
2. Incorporating suggestions

V. Judgment and Skill in Implementing Treatment

A. Planning the Operation

1. Reviewing literature, X-rays, other material
2. Planning approach and procedures

B. Making Necessary Preparations for Operating

1. Preparing and checking patient
2. Readying staff, operating room, supplies

C. Performing the Operation

1. Asking for confirmation of involved area
2. Knowing and observing anatomical principles
3. Using correct surgical procedures
4. Demonstrating dexterity or skill
5. Taking proper precautions
6. Attending to details
7. Persisting for maximum result

D. Modifying Operative Plans According to Situation

1. Deviating from pre-planned procedures
2. Improvising with implements and materials
3. Terminating operation when danger in continuing

E. Handling Operative Complications
  1. Recognizing complications
  2. Treating complications promptly and effectively

F. Instituting a Non-Operative Therapy Program
  1. Using appropriate methods and devices
  2. Applying methods and devices correctly

VI. Effectiveness in Treating Emergency Patients

A. Handling Patient
  1. Properly applying splints and other protective measures
  2. Handling and transporting carefully

B. Performing Emergency Treatment
  1. Determining location and extent of injuries
  2. Attending immediately to lifesaving procedures
  3. Treating most critical needs first
  4. Obtaining and organizing help

VII. Competence in Providing Continuing Care

A. Paying Attention Post-Operatively
  1. Administering suitable post-operative care
  2. Recognizing post-operative complications
  3. Adequately treating post-operative complications

B. Monitoring Patient's Progress
  1. Checking on effectiveness of therapy
  2. Reassessing, altering or repeating treatment

C. Providing Long-Term Care
  1. Arranging for rehabilitative care, socio-economic assistance
  2. Explaining and monitoring home and rehabilitative care

VIII. Effectiveness of Physician-Patient Relationship

A. Showing Concern and Consideration
  1. Taking personal interest
  2. Acting in discreet, tactful, dignified manner
  3. Avoiding needless alarm, discomfort, or embarrassment
  4. Speaking honestly to patient and family
  5. Persuading patient to undertake needed care, or only needed care

B. Relieving Anxiety of Patient and Family

1. Reassuring, supporting or calming
2. Explaining condition, treatment, prognosis or complication

IX. Accepting Responsibilities of a Physician

A. Accepting Responsibility for Welfare of Patient

1. Heeding the call for help
2. Devoting necessary time and effort
3. Meeting commitments
4. Insisting on primacy of patient welfare
5. Delegating responsibilities wisely
6. Adequately supervising residents and other staff

B. Recognizing Professional Capabilities and Limitations

1. Doing only what experience permits
2. Asking for help, advice or consultation
3. Following instructions and advice
4. Showing conviction and decisiveness
5. Accepting responsibility for own errors
6. Referring cases to other orthopedists and facilities

C. Relating Effectively to Other Medical Persons

1. Supporting the actions of other physicians
2. Maintaining open and honest communication
3. Helping other physicians
4. Relating in discreet, tactful manner
5. Respecting other physician's responsibility to his patient

D. Displaying General Medical Competence

1. Detecting, diagnosing, (treating) non-orthopedic disorders
2. Obtaining appropriate referrals
3. Preventing infection in hospital patients
4. Effectively keeping and following records

E. Manifesting Teaching, Intellectual and Scholarly Attitudes

1. Lecturing effectively
2. Guiding and supporting less experienced orthopedists
3. Encouraging and contributing to fruitful discussion
4. Contributing to medical knowledge
5. Developing own medical knowledge and skills

F. Accepting General Responsibilities to Profession and Community

1. Serving the profession
2. Serving the community
3. Maintaining personal and intellectual integrity

Appendix 2

A Taxonomy of Intellectual Processes *

The hierarchical ordering of this taxonomy is intended to imply that a certain degree of achievement at one or more of the lower levels is a necessary, though not a sufficient, condition for success at any higher level. It therefore follows that examinations which contain a large number of questions at the higher levels do not minimize the fundamental importance of basic information. Indeed they stress it, by assuring that the candidate cannot pass the examination unless he has the information at his command, and understands it sufficiently to make use of it in solving new problems.

LEVEL 1: RECALL

Items testing predominantly the RECALL of isolated information. This will include recognition of typical morphologic lesions and questions about specific facts, concepts, principles, processes, and theories. Whether or not it is explicitly so formulated, such a question will ordinarily be asking: "What is X?"

and

Items testing recognition of MEANING or implication. Such items will require the student to provide something more than a textbook or classroom answer, but do not require any significant degree of interpretation or application. Items of this type differ from those described in the above in that they will ask, for example, NOT "What is a blood test or what does it do?" but instead will ask: "Since a blood test does X, what does this mean that you can learn from it about Y?"

Illustration: The tensor fascia lata muscle is innervated by the:

a. inferior gluteal
b. femoral
c. superior gluteal
d. obturator

---

* Prepared by the Committee on Student Appraisal, University of Illinois, College of Medicine.

LEVEL 2: GENERALIZATION OR EXPLANATION

Items requiring the student to select a relevant GENERALIZATION to explain specific phenomena. Ordinarily items of this type will differ from Level 1 items in asking "Why is X true?" or "How do you explain X?" rather than in asking "Is X true?" or "What does X mean or imply?"

Illustration: The single most important step in the surgical technique of amputation is:

a. suturing of the fascia
b. bring a muscle pad over the bone end
c. accurate proportions between the anterior and posterior flap design
d. careful hemostasis and if necessary a drain for a short period postoperatively
e. accurate approximation of the flap edges

LEVEL 3: PROBLEM-SOLVING OF A FAMILIAR TYPE

Items requiring the student to make SIMPLE INTERPRETATIONS of DATA. Items of this type require the student to translate verbal, tabular, morphologic, or graphic data into another form (i.e., to read the data), or to make interpolations or extrapolations from the data.

and

Items requiring the student to APPLY a single principle or A STANDARD COMBINATION OF PRINCIPLES to a situation of a familiar type. In items of this type, while the specific content of the problem will be new to the student, the problem will involve a familiar pattern of attack.

Illustration: A 21 year old white male is involved in an auto accident sustaining a laceration of the face and an obvious closed mid third fracture of the left femur. The patient complains of chest pain and is short of breath. B/P 100/90, R 30, Pulse 100

What roentgenogram should be obtained? The patient's general condition is adequate to permit all indicated films to be made

a. skull films and x-rays of left femoral shaf

b. left femoral shaft, left hip in AP view, chest
c. left femoral shaft and chest
d. AP and lateral view of chest

LEVEL 4: PROBLEM-SOLVING OF AN UNFAMILIAR TYPE

Items requiring the ANALYSIS of data. Items of this type will require the student to recognize the constituent elements and relationships in a set of data, to judge their internal consistency and to comprehend the organizational principles involved.

and

Items requiring the student to APPLY A UNIQUE COMBINATION OF PRINCIPLES to solve a problem of a novel type. In items of this type, both the specific content and the character of the problem will be new to the student to the extent that solution will require a novel pattern of attack, not previously illustrated in classroom or textbook problems.

Illustration: (Note: The following item was preceded by a description of the presenting complaint, a brief history, a few questions, and additional data about the course of the disease process over the next two weeks.)

On the basis of this additional information, which of the following measures might provide the most useful information?

a. an electromyographic study of the abdominal musculature right and left
b. spinal tap with spinal fluid analysis
c. a complete blood count
d. additional x-ray studies of the thoracolumbar region
e. an erythrocyte sedimentation rate

LEVEL 5: EVALUATION

Items requiring the EVALUATION of a total situation. Items of this type may be based on a case report of the type prepared for the typical clinical-pathological conference, or a research report, or the presentation of a theory together with evidence, and will require the student to evaluate the total presentation.

Illustration: (Note: The following question was preceded by a description of the presenting complaint, a brief history, a few questions and then several sets of additional data including information on subsequent course.)

In the face of the present circumstances, which of the following procedures seems the most logical at this juncture?

a. immediate myelographic study
b. enforced recumbency on a turning frame with cephalopelvic traction
c. immediate laminectomy and decompression
d. electromyelographic study of the lower extremity musculature to determine the precise level of involvement

LEVEL 6: SYNTHESIS

Items requiring SYNTHESIS of a variety of elements of knowledge into an original and meaningful whole. Items of this type may be based on a clinical report which requires the student to develop a differential diagnosis or a therapeutic regimen. Alternatively, such questions may be based on a set of data which require the student to develop an original (to him) theory explaining the phenomena. Such items will involve the process of working with concepts and principles, and arranging and combining them in such a way as to constitute a pattern or structure not clearly there before.

Illustration: None

Appendix 3

WORKING PAPERS: TASK FORCE ON WRITTEN EXAMINATIONS

Task: To determine what kinds of competence are being measured in the written examinations currently in use.

Procedure: To classify each question in the January, 1964, Part II and May, 1964, Part I examinations, according to the kind of intellectual process the candidate is most likely to employ in answering the question.

Preparation: In advance of the meeting on December 27, each member of the study group should study the attached documents, and make a tentative classification of the questions appearing on pp. 12 to 19, and be prepared to suggest needed changes in the classification system.

DETERMINING WHAT A TEST MEASURES

We know from previous research on examinations, that some questions in a test can be answered by immediate recall of information. Others require the candidate to reason out the answer. It may, of course, be impossible to answer questions of this latter type unless one can recall certain basic information assumed by the questions, but the significant distinction is that it is impossible to answer the latter type of question exclusively on the basis of information recalled. The candidate must be able to go well beyond a process which relies on rote memory, in using the information at his command, to reason out the answer.

Among the questions which require the candidate to reason out the answer, the kind of reasoning process involved will vary with different kinds of questions. Once it is clear that a question cannot be answered from rote memory alone, then it is necessary to take a second step: i.e. to decide what kind of mental process one would ordinarily go through in order to answer the question. For example: Does it require the examinee to apply principles? To evaluate data? To analyze a problem? The following sections of this document list the various such processes to be used in the classification of the Board Examinations, together with illustrations of each category and a definition sufficient for the purposes of this classification. No attempt has been made to define categories which are philosophically and psychologically pure; instead the effort has been made to define categories that are sufficiently discrete to permit medical educators to identify the predominant characteristic of each question.

A few words of caution are necessary in using this classification:

1. In deciding first whether or not a question can be answered on the basis of immediate recognition and recall, it is necessary to look at the incorrect, as well as the correct, answers. By a process of elimination a candidate may be able to answer an apparently thought-testing question on the basis of recall only by excluding obviously wrong answers (and thereby coming to the correct answer without knowing or reasoning it out); he merely recalls that all the others are wrong.
2. In deciding whether or not a question can be answered on the basis of recall, it is necessary to consider both the average training program, standard references, typical experience and such other aids to learning as are normally available to a candidate. A man may "come" to a most impressive conclusion, but it may represent only the most superficial recall if it is of the type that he would ordinarily memorize from the standard references employed in his field.
3. In deciding whether or not a question can be answered on the basis of recall, it is necessary to try to imagine how the candidate would approach the question in an examination situa-

tion. For example, items that look like simple informational questions may actually be ones that candidates characteristicall *reason out.* With the amount of knowledge it is necessary to acquire in this field, it may be that physicians characteristically develop a system of thinking about an area which enables them to reconstruct the details through a reasoning, not a memory, process. Alternatively, it should be noted that questions related to case materials may involve simple recall and no reasoning, if the case description is so cut-and-dried that it represents a classical textbook description of symptoms or if some of the questions can be answered without specific reference to the case material.

To repeat: For purposes of this study, it is necessary to try to classify each question on the basis of how a candidate for certification *would approach it in an examination situation.*

4. You will note that the following classification (See Appendix 2) is arranged in an hierarchical system. If an item involves two or more levels (for instance, both recall and application) it should be classified at the *highest* level necessary to use in answering it.

Appendix 4: Observation Form and Instruction for Use

Sheet No.________
of ______ Sheets

Subject____________ Panel________ Examiners____________ + ____________ Candidate No.______ Observer____________

TIME		QUESTION	RESPONSE									EX	COMMENTS
			Process					Cites					
Note at beginning of each new subject, and more frequently if possible.	Visual Stimulus*	Specific Question (Record Each)	Recall	Prob. Solv.	Interp. Skill	Guess	Doesn't Know	Authority	Data	Experience	Demonstr.	Cues	NOTE any questions based on a concrete clinical situation, and types of stimulus, response, feedback, etc. not specifically provided for.

*KEY: M = Microscope; P = Photograph; S = Skeleton; X = X-Ray; O = Other

TIME	Visual Stimulus *	QUESTION	RESPONSE									EX	COMMENTS
Note at beginning of each new subject, and more frequently if possible.		Specific Question (Record Each)	Process					Cites					NOTE any questions based on a concrete clinical situation, and types of stimulus, response, feedback, etc. not specifically provided for.
			Recall	Prob. Solv.	Interp. Skill	Guess	Doesn't Know	Authority	Data	Experience	Demonstr.	Cues	

*KEY: M = Microscope; P = Photograph; S = Skeleton; X = X-Ray; O = Other

Examiners Comments:

Summary Notes:

GRADES		
1st Exam	2nd Exam	Final

INSTRUCTIONS TO OBSERVERS

Start each observation on a new sheet; use as many sheets per candidate as required. Be sure to record sheet number and total number of sheets for that candidate in upper right corner of front page of each sheet.

For each observation, fill in general information on top line of first sheet.

Clip together all sheets for each observation.

TIME:	In this column record time at which each new topic was introduced; in addition, record time of specific questions as frequently as possible.
VISUAL STIMULUS:	Whenever visual material is used as the basis for a question, record the appropriate SYMBOL (see key at bottom of form). Leave blank when no visual material is introduced as a part of the question.
SPECIFIC QUESTION:	Record EACH question with sufficient specificity to give a reader a clear indication of the nature of the question. (Each question in a series should be individually recorded and coded.)
RESPONSE PROCESS:	Check ONE of the columns labelled "Recall," "Problem-Solving," or "Interpretive Skill" to indicate the PREDOMINANT nature of the process revealed by the candidate's response. In general, only one of these three columns should be checked. However, if you are in real doubt put a question mark in a second column. When there is evidence that the candidate is guessing or doesn't know the answer, place a check in the appropriate column. The "Guess" and/or "Doesn't Know" columns may be checked whether or not there is a check in a "Recall," "Problem-Solving," or "Interpretive Skill" column. If some other process is involved, note in the Comment column.

OVER

[illegible]: Check the appropriate column or combination of columns whenever the candidate employs authority, data, his own experience, or a demonstration in answer to a question. If none of the foregoing is used, leave this group of columns blank. If the candidate supports or supplements his answer in some other way note in the Comment column.

EXAMINER CUES: Check this column whenever the examiner provides cues to lead the candidate in his answer. If the examiner scores the answer by indicating whether it is correct, or uses this occasion to teach, note under comments.

Check "Cues" only when the examiner rephrases his question, offers hints or suggestions or asks leading questions to assist the candidate; do NOT use it when the examiner merely says, "What else?" to mean: "Do you have anything to add to your answer?"

COMMENTS: Record any ADDITIONAL information that will assist in determining the nature of the examination or the setting. For example, types of stimulus material, responses or feedback not specifically provided for in the columns, distractions or assists from the second examiner, etc. Since the verbal column has been omitted from the stimulus material, also note in the Comment column any question that is based on a separate clinical problem about a specific patient. DO NOT so classify clinically oriented problems of a general nature.

At the end of a group of questions, draw a wavy line to separate questions dealing with different topics or cases.

At the conclusion of the examination record in the appropriate box at the end of the form, any comments of the examiner, your own summary notes and the candidate's grades.

Appendix 5

RESIDENT EVALUATION FORM

DO NOT WRITE
IN THIS SPACE

Col.No. Name of Resident ____________________________________

1 - 3 Identification No. ☐☐☐

4 - 6 Institution ____________________________________ Code ☐☐☐

7 - 9 Name of Rater ____________________________________ Code ☐☐☐

In filling out this form you are to rank the resident on each factor in terms of all the residents in orthopaedic surgery you have known during your career. You are to indicate your rankings by checking the appropriate box under each factor. In making these evaluations DO NOT take into account the resident's level of training. For example, a second year resident may have the potentiality to display outstanding surgical skills, but many fourth year residents might function AT THE PRESENT time at a higher level. He should be ranked lower than they are ranked on surgical skill. If you believe that you do not have sufficient information on the resident to evaluate a particular factor, check the appropriate box. Please write your name in the space above. All the information collected will be held strictly confidential and will not be used for any purpose other than research purposes.

- -

Col.
No.

Factor I: Ability to recall factual information concerning general medicine and orthopaedic surgery

This factor deals with the resident's comand of the factual information required of a practicing orthopaedist. Residents who score high are those who have a great deal of pertinent information at their "fingertips." Residents who score low are those who consistently display wide gaps in their knowledge. Residents can score well on this factor and low on Factor II below. They may recall a great deal of information, but have difficulty in integrating the information in solving problems in patient treatment and care.

10 I do not have sufficient information to judge. 1 ☐

RANKING

11-12

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest qurter		

Col. No. Factor II: Ability to use information to solve problems

This factor deals with the resident's effectiveness in using the information he has collected and recalled in solving problems in treatment and diagnosis.

13 I do not have sufficient information to judge. 1 ☐

RANKING

14-15

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

- -

Col. No. Factor III: Ability to gather clinical information

This factor deals with the resident's effectiveness in gathering clinical information. Is he generally thorough and discriminating, or does he fail to gather important information and in general is haphazard and inefficient in this factor?

6 I do not have sufficient information to judge. 1 ☐

RANKING

17-18

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

- -

Col. No. Factor IV: Judgment in deciding on appropriate treatment and care

This factor deals with the resident's ability to properly weigh the many factors involved in deciding on treatment and care, and to come to sound conclusions.

19 I do not have sufficient information to judge. 1 ☐

RANKING

0-21

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

- -

Col.
No.

Factor V: Skill in surgical procedures

This factor deals with the resident's manipulative skill in carrying out the procedures required of orthopaedists.

22 I do not have sufficient information to judge. 1 ☐

RANKING

23-24

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

Col.
No.

Factor VI: Relating effectively to patients

This factor deals with the resident's tact, consideration and skill in dealing with patients.

25 I do not have sufficient information to judge. 1 ☐

RANKING

26-27

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

Col.
No.

Factor VII: Relating effectively to colleagues and other medical personnel

This factor deals with how effectively the physician works as a member of a medical team, in asking advice, giving advice and showing tact and consideration.

28 I do not have sufficient information to judge. 1 ☐

RANKING

29-30

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

Col. No.

Factor VIII: Demonstrating the moral and ethical standards required of a physician

This factor deals with the resident's standards in terms of his concern for patients, his financial dealings, and his contacts with other physicians and society in general.

31 I do not have sufficient information to judge. ☐ 1

RANKING

32-33

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

- -

Col. No.

Factor IX: Overall competence as an orthopaedic surgeon

34 I do not have sufficient information to judge. ☐ 1

RANKING

35-36

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

- -
- -

37-40 Date completed____________________________

Appendix 6

Examination Number__________

Resident Evaluation Form

1967 ORTHOPAEDIC IN-TRAINING EXAMINATION

TO CHIEF OF SERVICE

Please complete one of these forms for each resident who will take the 1967 Orthopaedic In-Training Examination and return these with the examination answer sheets to:

American Academy of Orthopaedic Surgeons
29 East Madison Street
Chicago, Illinois 60602

This information will be used for statistical purposes only and will be kept strictly confidential.

Do Not Write
In This Space

Col. No. 9-29 RESIDENT'S NAME __________

Col. No. 30 YEAR in TRAINING __________

RANKING

			Lower Quarter	Lower Middle Quarter	Upper Middle Quarter	Upper Quarter
Col. No. 31	Factor 1:	KNOWLEDGE OF CLINICAL ORTHOPAEDICS	☐ 1 ☐ 2	☐ 3 ☐ 4	☐ 5 ☐ 6	☐ 7 ☐ 8
Col. No. 32	Factor 2:	KNOWLEDGE OF BASIC SCIENCES AS RELATED TO ORTHOPAEDICS	☐ 1 ☐ 2	☐ 3 ☐ 4	☐ 5 ☐ 6	☐ 7 ☐ 8
Col. No. 33	Factor 3:	ABILITY TO GATHER CLINICAL INFORMATION	☐ 1 ☐ 2	☐ 3 ☐ 4	☐ 5 ☐ 6	☐ 7 ☐ 8
Col. No. 34	Factor 4:	ABILITY TO USE INFORMATION TO SOLVE PROBLEMS	☐ 1 ☐ 2	☐ 3 ☐ 4	☐ 5 ☐ 6	☐ 7 ☐ 8
Col. No. 35	Factor 5:	JUDGMENT IN DECIDING APPROPRIATE TREATMENT AND CARE	☐ 1 ☐ 2	☐ 3 ☐ 4	☐ 5 ☐ 6	☐ 7 ☐ 8
Col. No. 36	Factor 6:	SKILL IN SURGICAL PROCEDURES	☐ 1 ☐ 2	☐ 3 ☐ 4	☐ 5 ☐ 6	☐ 7 ☐ 8
Col. No. 37	Factor 7:	RELATING EFFECTIVELY to PATIENTS	☐ 1 ☐ 2	☐ 3 ☐ 4	☐ 5 ☐ 6	☐ 7 ☐ 8
Col. No. 38	Factor 8:	RELATING EFFECTIVELY TO COLLEAGUES AND OTHER MEDICAL PERSONNEL	☐ 1 ☐ 2	☐ 3 ☐ 4	☐ 5 ☐ 6	☐ 7 ☐ 8
Col. No. 39	Factor 9:	DEMONSTRATING THE MORAL AND ETHICAL STANDARDS REQUIRED OF A PHYSICIAN	☐ 1 ☐ 2	☐ 3 ☐ 4	☐ 5 ☐ 6	☐ 7 ☐ 8
Col. No. 40	Factor 10:	OVER-ALL COMPETENCE AS AN ORTHOPAEDIC RESIDENT	☐ 1 ☐ 2	☐ 3 ☐ 4	☐ 5 ☐ 6	☐ 7 ☐ 8

Appendix 7

AMERICAN BOARD OF ORTHOPAEDIC SURGERY

29 East Madison Street
Chicago, Illinois 60602

CANDIDATE EVALUATION FORM

INSTRUCTIONS: The physician below has applied for entrance to the Certification Examination of the American Board of Orthopaedic Surgery. In reviewing his application, the Board would like to have some information on his capabilities in each of the areas of competence listed on the following pages. For each area, a description has been prepared of the effective and ineffective physician. Please indicate where you believe the candidate fits in this continuum by drawing a vertical line across some point on the line below each description of the factor.

DO NOT WRITE IN THIS SPACE

Name of Candidate ______________________________
Last, First

1 - 10 Identification Number ______________________________

11 - 15 Date form filled in ______________________________

Prepared with the assistance of
Center for the Study of Medical Education
University of Illinois College of Medicine

DO NOT
WRITE
IN THIS
SPACE

Please fill out the following information about yourself.

Name of Rater: ______________________________
Last First

16 - 25 Rater's ID No. ______________ (To be filled out by CSME)

26 - 29 Rater's Institution ______________________________

Your relationship to Candidate (Check as many as apply)

30 ☐ Chief of Training at institution where he trained
31 ☐ Other full-time orthopaedist at institution where he trained
32 ☐ Other full-time non-orthopaedic physician at institution where he trained
33 ☐ Fellow orthopaedic resident at institution where he trained
34 ☐ Fellow non-orthopaedic resident at institution where he trained
35 ☐ Attending orthopaedist at hospital where he trained
36 ☐ Attending non-orthopaedic physician at hospital where he trained
37 ☐ Orthopaedic colleague in community where he practices
38 ☐ Non-orthopaedic colleague in community where he practices
39 ☐ Other (Specify) ______________________________

40 **Period of acquaintanceship with candidate:**

☐ 1 **0 - 6 mos.** ☐ 2 **6 - 12 mos.**

☐ 3 **1 - 3 yrs.** ☐ 4 **3 - 5 yrs.** ☐ 5 **over 5 yrs.**

41 **Familiarity with candidate's practice:**

☐ 1 **Not familiar** ☐ 2 **Slightly familiar**

☐ 3 **Moderately familiar** ☐ 4 **Very familiar**

Col. No.

Factor 1. INFORMATION GATHERING

This factor is concerned with the Candidate's willingness, ability and skill in gathering information necessary for diagnosis.

The INEFFECTIVE Candidate limits his interview and physical examination to the area of complaint and fails to pursue alternative hypotheses.

He frequently uses therapy to substantiate clinical impressions.

The EFFECTIVE Candidate routinely takes a comprehensive initial history and physical examination. He records the information received in a systematic fashion, and pays careful attention to progress notes.

He is aware of information other than the medical and indicates this by initiating further procedures and questions.

42 - 43

01	02	03	04	05	06	07	08	09	10	11	12
Poor			Marginal			Good			Excellent		

44 ☐ Insufficient information to judge

Factor 2. PROBLEM-SOLVING

This factor is concerned with the Candidate's ability and skill in using information gained to develop a diagnosis and support clinical activity.

The INEFFECTIVE Candidate has an incomplete comprehension of the implications of the data he has collected.

He is unable to interpret unexpected results and often ignores them.

He makes decisions on the basis of experience, disregarding the context in which that experience was gained.

His thinking is rigid and unimaginative, impeding his recognition of associated problems.

The EFFECTIVE Candidate realizes the importance of unexpected findings and seeks to determine their implications.

He understands the nature of probability and uses this to illuminate his experience.

He takes all the data into account before reaching a decision, and routinely tests alternative hypotheses.

45 - 46

01	02	03	04	05	06	07	08	09	10	11	12
Poor			Marginal			Good			Excellent		

47 ☐ Insufficient information to judge

Col. No.

Factor 3. CLINICAL JUDGMENT

This factor is concerned with the Candidate's ability to use sound judgment in planning for and carrying out treatment.

The INEFFECTIVE Candidate is overly concerned with treatment techniques at the expense of overall goals.

He often delegates pre- and post-operative care to others.

He plans treatment without sufficient familiarity with the procedures he selects.

His treatment choice is rigid--using a set formula for treating each clinical problem or using a favorite technique when more effective ones are available.

The EFFECTIVE Candidate is familiar with the uses and limitations of the procedures he attempts. He recognizes his own capabilities and uses procedures which correspond to them.

He considers simple procedures first.

His clinical judgment encompasses information beyond the pathologic.

He demonstrates regard for patients' needs, desires and life conditions.

He is flexible enough to modify his treatment plans when the situation warrants doing so.

48-49

01	02	03	04	05	06	07	08	09	10	11	12
Poor			Marginal			Good			Excellent		

50 ☐ Insufficient information to judge

Factor 4. SURGICAL TECHNIQUE

This factor is concerned with the Candidate's ability and skill in carrying out operative procedures.

The INEFFECTIVE Candidate has insufficient skill for the procedures he attempts.

His overall handling of instruments and tissue lacks finesse.

His operating time is often prolonged through unfamiliarity with procedures or inadequate planning.

He takes unnecessary operative risks or terminates operation before maximum results are achieved.

The EFFECTIVE Candidate handles tissues gently, uses careful haemostases, and makes a proper and adequate exposure of the operating field.

He carefully attends to details such as sterilization of instruments and proper choice of same.

He makes proper application of fixation devices or prosthesis and makes proper closure of wounds.

He carefully monitors his patient during operative procedure.

He applies appropriate dressings, splints and casts.

51-52

01	02	03	04	05	06	07	08	09	10	11	12
Poor			Marginal			Good			Excellent		

53 ☐ Insufficient information to judge

Col. No.

Factor 5. RELATING TO PATIENT

This factor is concerned with the Candidate's effectiveness in working with patients.

The INEFFECTIVE Candidate does not communicate with his patients, either through aloofness, indifference or the pressure of time.

He has difficulty understanding patient needs.

He is unable to evoke patient confidence, tending even to alarm them.

He reacts negatively to hostility or other emotional displays.

The EFFECTIVE Candidate's man elicits patient confidence and cooperation and relieves anxiety.

He is interested in his patient's well-being and demonstrates this without becoming emotionally involved.

He is honest with the patient and his family.

Patients like him and readily feel they can ask questions and discuss problems with him.

54-55	01	02	03	04	05	06	07	08	09	10	11	12
	Poor			Marginal			Good			Excellent		

56 ☐ Insufficient information to judge

Factor 6. CONTINUING RESPONSIBILITY

This factor is concerned with the Candidate's willingness to accept the responsibility for long-term patient care.

The INEFFECTIVE Candidate either loses interest after initial treatment or does not take the time for adequate follow-up.

He becomes discouraged with slow progress and cannot cope with a poor prognosis. He is unable to communicate realistic expectations to the patient.

His utilization of support personnel is either inadequate or he expects assistance beyond their capabilities and training.

The EFFECTIVE Candidate is able and willing to work with the patient to achieve maximum rehabilitation. He motivates the patient to strive for his own rehabilitation.

He monitors patients' progress, altering therapy or treatment as indicated.

He understands the roles of various allied health professions and makes maximum use of their assistance.

He maintains a positive and persistent attitude toward recovery.

57-58	01	02	03	04	05	06	07	08	09	10	11	12
	Poor			Marginal			Good			Excellent		

59 ☐ Insufficient information to judge

Col. No.

Factor 7. EMERGENCY CARE

This factor is concerned with the Candidate's ability to act effectively in emergency situations, in the operating theatre or the emergency room.

The INEFFECTIVE Candidate panics easily and makes inappropriate use of time available.

He becomes confused under pressure and has difficulty establishing priorities. He is unable to delegate aspects of care to others.

He is careless about applying protective measures.

He is unable to make decisions alone.

The EFFECTIVE Candidate quickly assesses the situation, pays attention to lifesaving procedures and demonstrates understanding of triage concepts.

He is able to obtain and organize assistance of others.

He is able and willing to make decisions alone if necessary.

He is aware of the consequences of delay.

60-61

01	02	03	04	05	06	07	08	09	10	11	12
Poor			Marginal			Good			Excellent		

62 ☐ Insufficient information to judge

Factor 8. RELATING TO COLLEAGUES

This factor is concerned with the Candidate's ability to work effectively with his colleagues and other members of the health team.

The INEFFECTIVE Candidate has difficulty relating to others and lacks the ability either to give or take instruction gracefully.

He tends to be tactless and inconsiderate and does not evoke the confidence and cooperation of those with whom he works.

He habitually gives unsolicited advice, and in an offensive manner.

He is unwilling to make referrals or seek consultation and fails to support his colleagues in their contacts with his patients.

The EFFECTIVE Candidate relates well to others and communicates easily, working well in a team situation.

He seeks consultation when appropriate and respects others' views.

He demonstrates self-control.

He gives credit to others for their contributions and creates an atmosphere of working together--not working for.

63-64

01	02	03	04	05	06	07	08	09	10	11	12
Poor			Marginal			Good			Excellent		

65 ☐ Insufficient information to judge

4

Col. No.

Factor 9. MORAL AND ETHICAL VALUES

This factor is concerned with the Candidate's attitudes and standards as an individual.

The INEFFECTIVE Candidate attempts to cover up his errors.

He is frequently absent from assigned duty or unavailable when needed.

He has unethical contacts with non-medical professions and allows his personal finances to unduly influence treatment.

He discusses medical mismanagement with patients.

The EFFECTIVE Candidate's conduct reflects kindness, respect, honesty and humility.

He reports facts accurately, including his own errors.

He respects the confidences of colleagues and patients.

He places patient care above personal considerations.

He respects the property of others.

He recognizes his own professional capabilities and limitations.

66-67

01	02	03	04	05	06	07	08	09	10	11	12
Poor			Marginal			Good			Excellent		

68 ☐ Insufficient information to judge

Factor 10. OVERALL COMPETENCE

This factor is concerned with your judgment of the Candidate's overall competence as an orthopaedic surgeon, taking into account Factors 1 through 9.

69-70

01	02	03	04	05	06	07	08	09	10	11	12
Poor			Marginal			Good			Excellent		

71 ☐ Insufficient information to judge

Col. No.

ADDITIONAL COMMENTS

You are encouraged to comment further on any or all of the Factors in this Evaluation Form or to bring additional information to the attention of the Eligibility Committee. Please use the space below for these comments.

72

Appendix 8

Distribution of Ratings, Candidate Evaluation form, for 1968 Final Certification Examination

Rating	Factors 1		2		3		4		5		6		7		8		9		10	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
12 E x	140	8.8	108	6.9	111	7.1	123	7.8	182	11.6	185	11.8	164	10.4	194	12.3	385	24.5	108	6.9
11 c e			200	12.7	208	13.2	202	12.8	252	16.0	239	15.2	240	15.2	238	15.1	325	20.6	242	15.4
10 l	516	32.7	309	19.6	331	21.0	363	23.1	354	22.5	369	23.4	385	24.5	368	23.4	349	22.2	356	22.6
9 G			302	19.2	316	20.0	312	19.8	242	15.4	268	17.0	296	18.8	248	15.8	193	12.3	309	19.6
8 o o	562	36.7	269	17.1	249	15.8	264	16.8	221	14.0	222	14.1	186	11.8	214	13.6	136	8.6	254	16.1
7 d			222	14.1	199	12.6	155	9.8	153	9.7	129	8.2	114	7.2	149	9.5	86	5.5	174	11.0
6 M	271	17.2	77	4.9	78	5.0	53	3.4	81	5.1	48	3.0	31	2.0	67	4.3	26	1.7	56	3.6
5 a r			26	1.7	22	1.4	11	0.7	18	1.1	20	1.3	12	0.8	21	1.3	14	0.9	15	1.0
4 g	31	2.0	6	0.4	7	0.4	11	0.7	11	0.7	9	0.6	6	0.4	14	0.9	7	0.4	5	0.3
3 P			8	0.5	5	0.3	6	0.4	7	0.4	5	0.3	2	0.1	8	0.5	3	0.2	5	0.3
2 o o	6	4	4	0.3	3	0.2	2	0.1	2	0.1	0	0	0	0	5	0.3	5	0.3	0	
1 r			1	0.1	0	0.0	0	0.0	1	0.1	2	0.1	0		4	0.3	5	0.3	0	
Could not rate	48	3.0	42	2.6	45	2.9	72	4.6	48	3.0	78	5.0	138	8.8	44	2.8	40	2.5	48	3.0

* The Computer printout for this factor classified the ratings into only 6 categories, so that the 140, 12's include 11's as well, etc.

Titles of Factors

1. Information Gathering
2. Problem Solving
3. Clinical Judgment
4. Surgical Technique
5. Relating to Patient
6. Continuing Responsibility
7. Emergency Care
8. Relating to Colleagues
9. Moral and Ethical Values
10. Overall Competence

See Appendix 6 for a description of these Factors

[illegible] 197
Appendix [illegible]

[illegible]

[illegible]

DO NOT WRITE
IN THIS SPACE

1 - 20 Name of Resident ______________________________
(Last) (First) (Initial)

21 - 24 Identification No. ☐ ☐ ☐ ☐

25 Year of hospital training after internship ________ ☐

26 - 28 Surgical procedure ________ ☐ ☐ ☐

29 - 30 Anesthesia ________ ☐ ☐

31 - 34 Institution ________ ☐ ☐ ☐ ☐

35 - 37 Examiner ________ ☐ ☐ ☐

The purpose of filling out this form is to assist in determining the competence of a resident in performing basic surgical procedures in an actual operating environment. The evaluation is to be made by directly observing a surgical procedure by a resident in his accustomed environment. The rating form is to be filled out promptly following the procedure by an observer who will be scrubbed with the team as second surgical assistant performing only those tasks delegated by the surgeon. The rater will mark the appropriate rating in each section, and record comments in each section in which the resident demonstrated unusually outstanding or poor performance.

Factor I. Initial preparation for surgery

Did surgeon reconfirm procedure with patient before surgery?
Was surgeon properly gowned? (Cap over hair, mask tight, careful timed scrub of hands, gets into gown and gloves properly.)
Did he review procedure and position with anesthesia staff? Was the position of patient appropriate for procedure. (Surgery, tourniquet, X-rays, bone graft, etc.)
Did he know the names of nursing and medical staff?
Was preparation of patient's skin adequate? (Check area scrubbed, technique of applications, method of discarding sponges, etc.)
Was draping satisfactory and appropriate to procedures?
Did he prevent contamination by others?

4 ☐ Excellent
3 ☐ Good
2 ☐ Adequate
1 ☐ Poor

Comments

[illegible]

[illegible]?
[illegible]?
[illegible]?
Did [illegible] position (surgeon and instrument [illegible])?
Did he [illegible] the patient with [illegible] position [illegible]?
Was the placement and choice (if possible) of skin incision satisfactory?
Did the skin incision have adequate length?
Were the wound edges protected and skin instruments separated?
Did incision allow development of deep exposure?

4 ☐ Excellent
3 ☐ Good
2 ☐ Adequate
1 ☐ Poor

Comments

39

Factor III. Management of Tissues

Was hemostasis effective? (Precision in application of hemostats)?
Was cautery with minimal tissue damage, adequate pressure and time with use of sponges?
If tourniquet were used was venous congestion lessened?
Was continuous attention given to gentle handling of tissues?
Was exposure satisfactory without undue stretching or compression of tissues? Were retractors repositioned and changed with progress of procedure?
Were scalpel, scissors, dissecting probe used?
Was dissection guided by anatomy?
Was care taken for particular areas of dissection? (ex: protection of nerves, separation of vessels, etc.)
Was technique appropriate and adequate?
  (a) Efficient manipulation of fractures or bone fragments
  (b) Adequate release in transfer of tendon, nerve
  (c) Sufficient release in late correction of fracture and contractures
  (d) Were needs anticipated?
Was suture material specific to requirements?
Was tissue taken for pathology adequate and identified?

4 ☐ Excellent
3 ☐ Good
2 ☐ Adequate
1 ☐ Poor

Comments

[illegible]

[illegible]

[illegible]

Was [illegible]?

Internal fixation:

(a) Was appliance of adequate length, strength, etc.?
(b) Was fixation [illegible] applied?
(c) Were adjacent tissues protected during procedure?
(d) Were needs anticipated?

Was good technique followed regarding protection of stability?
Were instruments and equipment protected (scratches, bending, metal compatibility)?

41

4 ☐ Excellent
3 ☐ Good
2 ☐ Adequate
1 ☐ Poor

Comments

Factor V. Management of team personnel

Did all members have responsibility and did they continue to give intelligent aid to surgeon throughout procedure?
Was contamination prevented?
Was team advised by surgeon of his findings and progress of surgery?
Did all see wound?
Was distracting conversation and noise avoided?
Did surgeon request suggestions from team?
Did surgeon lose his temper?
Were all team members instructed in duties in emergency situations?
Did assistants aid surgical exposure?
Was some of surgical procedure shared?
Did surgeon thank nurses and others assisting?

42

4 ☐ Excellent
3 ☐ Good
2 ☐ Adequate
1 ☐ Poor

Comments

[illegible] Factor VI. [illegible]

[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
Were splints, plaster, special bed traction ready at proper time?
Did surgeon complete procedure including final dressing?

43
4 ☐ Excellent
3 ☐ Good
2 ☐ Adequate
1 ☐ Poor

Comments

Factor VII. Surgical management

Was procedure unduly rapid or prolonged?
Was initial purpose and plan carried out?
Was consideration given to general status of patient throughout procedure. Were supportive procedures or medications given?
Was blood loss consistent with good management?
Did surgeon demonstrate emotional control?
Was surgeon able to maintain interest and support of operating team?
Was adequate tissue requested?
Was X-ray control of procedure satisfactory?
Was prognosis given in realistic terms?

44
4 ☐ Excellent
3 ☐ Good
2 ☐ Adequate
1 ☐ Poor

Comments

Factor VIII. Overall competence in surgical skill

This factor involves your overall impression of the resident's surgical ability taking into account all seven factors above.

45 - 46
☐ 01 ☐ 02 ☐ 03 ☐ 04 ☐ 05 ☐ 06 ☐ 07 ☐ 08 ☐ 09 ☐ 10 ☐ 11 ☐ 12
Poor Adequate Good Excellent

-50 Date form was filled out ____________

Additional Comments

Appendix 10

ORAL EXAMINATION

AMERICAN BOARD OF ORTHOPAEDIC SURGERY

RATING FORM FOR USE WITH "PATIENT" INTERVIEWS

Candidate's Examination Number: ____________________ (Cols. 1 - 3)

Examiner's Name: ____________________ (Cols. 4 - 5)

Date: ____________________ (Cols. 6 - 7)

Starting Time: ____________________ (Cols. 8 - 11)

Prepared with the assistance of
CENTER FOR THE STUDY OF MEDICAL EDUCATION
UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE

Col. No.

RATING OF DIAGNOSTIC INTERVIEW

12 - 13 Diagnostic Case No. ______

Factor I: Ability to elicit an adequate amount of pertinent information

Weight 4

(The candidate should ask most of the indicated questions; other questions should be appropriate to the diagnosis.)

14 - 15	01	02	03	04	05	06	07	08	09	10	11	12
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
		Poor			Adequate			Good			Excellent	

Factor II: Ability to communicate with the patient

Weight 1

(Did he use appropriate vocabulary, use concepts familiar to the patient, and allow the patient to narrate parts of the history?)

16 - 17	01	02	03	04	05	06	07	08	09	10	11	12
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
		Poor			Adequate			Good			Excellent	

Factor III: Efficiency in gathering data

Weight 1

(Did he ask relevant and necessary questions, and avoid the time waste of exploring remote diagnoses which prevent an adequate examination of the pertinent facts?)

18 - 19	01	02	03	04	05	06	07	08	09	10	11	12
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
		Poor			Adequate			Good			Excellent	

Col. No.

RATING OF DIAGNOSTIC INTERVIEW (Continued)

Factor IV:
Ability to arrive at a diagnosis and present logical reasons for it
Weight 4

(Did he fail to consider all the pertinent facts he uncovered, make errors in relating or interpreting facts, or make errors in weighing the facts at hand?)

20 - 21 01 ☐ 02 ☐ 03 ☐ Poor | 04 ☐ 05 ☐ 06 ☐ Adequate | 07 ☐ 08 ☐ 09 ☐ Good | 10 ☐ 11 ☐ 12 ☐ Excellent

---

Factor V: Overall evaluation of Diagnostic Interview

22 - 23 01 ☐ 02 ☐ 03 ☐ Poor | 04 ☐ 05 ☐ 06 ☐ Adequate | 07 ☐ 08 ☐ 09 ☐ Good | 10 ☐ 11 ☐ 12 ☐ Excellent

---

24 Your role: ☐ "patient" 1 ☐ rater only 2

---

25 Comments:

The Candidate was difficult to evaluate because:

26 ☐ He spoke slowly

27 ☐ He spoke rapidly

28 ☐ He did not speak English well

29 ☐ He seemed excessively nervous

30 ☐ He seemed confused about the procedure

31 ☐ Other ____________________

32 I did not find the Candidate difficult to evaluate ☐

Col. No.	**RATING OF PROPOSED TREATMENT INTERVIEW**
33 - 34	Proposed Treatment Case No.__________
	Factor I: Effectiveness of the candidate's statements Weight 6 (Did he give too little information, oversimplify, indicate undue pessimism or optimism, overwhelm the patient with excessive detail or use inappropriate vocabulary?)
35 - 36	01 ☐ 02 ☐ 03 ☐ Poor 04 ☐ 05 ☐ 06 ☐ Adequate 07 ☐ 08 ☐ 09 ☐ Good 10 ☐ 11 ☐ 12 ☐ Excellent
	Factor II: Effectiveness of the candidate's manner Weight 2 (Was the manner in which the physician dealt with the "patient" one which would genuinely convince the patient that the physician is interested in his welfare?)
7 - 38	01 ☐ 02 ☐ 03 ☐ Poor 04 ☐ 05 ☐ 06 ☐ Adequate 07 ☐ 08 ☐ 09 ☐ Good 10 ☐ 11 ☐ 12 ☐ Excellent
	Factor III: Efficiency of the interview in terms of the interaction between patient and physician (Did the physician present the required information to the patient in a clear-cut efficient fashion?)
39 - 40	01 ☐ 02 ☐ 03 ☐ Poor 04 ☐ 05 ☐ 06 ☐ Adequate 07 ☐ 08 ☐ 09 ☐ Good 10 ☐ 11 ☐ 12 ☐ Excellent
	Factor IV: Overall evaluation of the Proposed Treatment Interview
41 - 42	01 ☐ 02 ☐ 03 ☐ Poor 04 ☐ 05 ☐ 06 ☐ Adequate 07 ☐ 08 ☐ 09 ☐ Good 10 ☐ 11 ☐ 12 ☐ Excellent
.3	Your role: ☐ "patient" 1 ☐ rater only 2
44	Comments:

Appendix 11

ORAL EXAMINATION

AMERICAN BOARD OF ORTHOPAEDIC SURGERY

RATING FORM
FOR
SIMULATED PATIENT MANAGEMENT CONFERENCE

Candidate's Number	Examiner's Overall Rating	Examiner's Converted Score	Combined Converted Score
A ________ (Col.1 - 3)	________ (Col.4 - 5)	________ (Col.6 - 7)	________ (Col.8 - 9)
B ________ (Col.1 - 3)	________ (Col.4 - 5)	________ (Col.6 - 7)	________ (Col.8 - 9)
C ________ (Col.1 - 3)	________ (Col.4 - 5)	________ (Col.6 - 7)	________ (Col.8 - 9)
D ________ (Col.1 - 3)	________ (Col.4 - 5)	________ (Col.6 - 7)	________ (Col.8 - 9)
E ________ (Col.1 - 3)	________ (Col.4 - 5)	________ (Col.6 - 7)	________ (Col.8 - 9)

Examiner's Number ____________________ (Col. 10 - 12)

Date: ____________________ (Col. 13 - 15)

Starting Time ____________________ (Col. 16 - 18)

Case Description Numbers ________ (Col. 19-20) and ________ (Col. 21-22)

Prepared with the assistance of the
Center for the Study of Medical Education
University of Illinois, College of Medicine

DESCRIPTION OF LEVELS

OF PERFORMANCE

---

0. **ERROR:** Reserve this rating for errors of fact or concept that can best be judged primarily in terms of content.

I. **HINDERS GROUP:** Distracts group with inappropriate, irrelevant or illogical suggestions, provokes group with antagonistic, argumentative statements that reflect or invite hostility; monopolizes discussion with repetitive insistence on own point of view.

II. **IS PASSIVE, NON-FACILITATIVE:** Shows passive acceptance or rejection; censures, complies, merely repeats or ratifies suggestions of others, makes personal, private, idiosyncratic comments or expresses ideas ineffectively or unclearly.

III. **CLARIFIES AND PROVIDES CONSTRUCTIVE SUGGESTIONS:** Clarifies issues; clearly and effectively presents useful suggestions, evaluations or pertinent information; amplifies on suggestions presented by others; relieves tension and promotes group consensus.

IV. **ORGANIZES, INTEGRATES, GREATLY FACILITATES:** Provides orientation or helps reorient group; assists others to participate; analyzes, summarizes, synthesizes discussion; reconciles differences and integrates ideas to achieve group solution.

OVERALL RATING

Candidates who say little or nothing should be rated poor; those for whom most of the tallies are in levels O, I and II should be rated no higher than poor; good or adequate ratings depend on the distribution of tallies between levels II and III; excellent rating requires that the majority of tallies be in levels III and IV.

RATING SCALE

INSTRUCTIONS: Place a tally mark in the appropriate box for EACH statement EACH candidate makes.

LEVEL OF STATEMENT	CANDIDATE				
	A	B	C	D	E
0. ERROR (Col. No.)	(23-24)	(23-24)	(23-24)	(23-24)	(23-24)
I. HINDRANCE (Col. No.)	(25-26)	(25-26)	(25-26)	(25-26)	(25-26)
II. PASSIVITY (Col. No.)	(27-28)	(27-28)	(27-28)	(27-28)	(27-28)
III. CLARIFICATION (Col. No.)	(29-30)	(29-30)	(29-30)	(29-30)	(29-30)
IV. INTEGRATION FACILITATION (Col. No.)	(31-32)	(31-32)	(31-32)	(31-32)	(31-32)

OVERALL RATING

INSTRUCTIONS: Write a figure from 0 - 12 to indicate your overall rating of each candidate

NOTE:

3 = Poor
4 - 6 = Adequate
7 - 9 = Good
10 - 12 Excellent

CANDIDATE				
A	B	C	D	E
(33-34)	(33-34)	(33-34)	(33-34)	(33-34)

COMMENTS

Appendix 12

ORAL EXAMINATION

SIMULATED PATIENT MANAGEMENT CONFERENCE

Part II: January 1966

RATING FORM
FOR
SIMULATED PATIENT MANAGEMENT CONFERENCE

Examiner's Number________________ (Col. 1 - 3)

Date: January________________ 1966 (Col. 4 - 5)

Starting Time________________ (Col. 6 - 9)

Case Description Numbers________ (Col. 10 - 11) and ________ (Col. 12 - 13)

Candidates Numbers

A________________ (Col. 14 - 15)

B________________ (Col. 16 - 17)

C________________ (Col. 18 - 19)

D________________ (Col. 20 - 21)

E________________ (Col. 22 - 23)

Prepared with the assistance of
Center for the Study of Medical Education
University of Illinois, College of Medicine

Factor I: Individual achievement

This factor deals with the quality of the candidates' discussion. It does not deal with their effectiveness as participants in a group.

Candidates who score HIGH are those who present solutions to the problem, effectively amplify on solutions presented by others, and express their ideas clearly, logically and effectively.

Candidates who score LOW are those who present few ideas of their own, i.e., who merely sit back and ratify the ideas of others, or who make inappropriate (irrelevant or unwise) recommendations or who express their ideas in an unclear, illogical fashion.

Col. No.		Poor			Adequate			Good			Excellent		
24 - 27	A	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07	☐ 08	☐ 09	☐ 10	☐ 11	☐ 12
28 - 31	B	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07	☐ 08	☐ 09	☐ 10	☐ 11	☐ 12
32 - 35	C	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07	☐ 08	☐ 09	☐ 10	☐ 11	☐ 12
36 - 39	D	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07	☐ 08	☐ 09	☐ 10	☐ 11	☐ 12
40 - 43	E	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07	☐ 08	☐ 09	☐ 10	☐ 11	☐ 12

Factor II: Ability to assist the group to reach its goals

This factor deals with the effectiveness of the candidates' participation as members of a group which is charged with the responsibility of achieving some objective.

Candidates who score HIGH are those who assist others to participate, summarize what others say, attempt to clarify issues and to reconcile differences of opinion in order to reach agreement and in general assist the group to reach some consensus in the allotted time.

Candidates who rank LOW are those whose statements impede the group in effectively exploring a topic and arriving at a consensus; they may contribute nothing or their contribution may be disruptive and divisive.

As is the case in Factor I, candidates who say very little or talk about irrelevancies should score low in this factor. Candidates who score high in Factor I can nevertheless score low in Factor II if their presentation so monopolizes the discussion as to impede achievement of the group's objectives.

	Poor			Adequate			Good			Excellent		
A	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07	☐ 08	☐ 09	☐ 10	☐ 11	☐ 12
B	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07	☐ 08	☐ 09	☐ 10	☐ 11	☐ 12
C	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07	☐ 08	☐ 09	☐ 10	☐ 11	☐ 12
D	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07	☐ 08	☐ 09	☐ 10	☐ 11	☐ 12
E	☐ 01	☐ 02	☐ 03	☐ 04	☐ 05	☐ 06	☐ 07	☐ 08	☐ 09	☐ 10	☐ 11	☐ 12

Factor III. Effective conduct as member of a group

This factor deals with how well the candidates work with colleagues in solving mutual problems and in handling professional differences. It differs from Factor II in that Factor II deals with what is said, but Factor III deals with how it is said.

Candidates who score HIGH are those who are able to accept disagreement without becoming upset, who refrain from sarcastic comments, avoid interrupting even when they obviously have something important to say, and in general give the impression that they welcome the participation of others.

Candidates who score LOW are those who have difficulty controlling their emotions, interrupt to an undue extent and in general show little concern for the feelings and ideas of others in presenting their statements.

Candidates can score high on Factors I and II and low on Factor III because they may present good statements and work hard to get the group to arrive at a consensus, but, in the process of doing so, they may antagonize others and cut off discussion abruptly. They would then be limiting participation and thus reducing the effectiveness of the group.

Col. No.

44 - 47 A ☐ 01 ☐ 02 ☐ 03 ☐ 04 ☐ 05 ☐ 06 ☐ 07 ☐ 08 ☐ 09 ☐ 10 ☐ 11 ☐ 12
Poor Adequate Good Excellent

48 - 51 B ☐ 01 ☐ 02 ☐ 03 ☐ 04 ☐ 05 ☐ 06 ☐ 07 ☐ 08 ☐ 09 ☐ 10 ☐ 11 ☐ 12
Poor Adequate Good Excellent

52 - 55 C ☐ 01 ☐ 02 ☐ 03 ☐ 04 ☐ 05 ☐ 06 ☐ 07 ☐ 08 ☐ 09 ☐ 10 ☐ 11 ☐ 12
Poor Adequate Good Excellent

56 - 59 D ☐ 01 ☐ 02 ☐ 03 ☐ 04 ☐ 05 ☐ 06 ☐ 07 ☐ 08 ☐ 09 ☐ 10 ☐ 11 ☐ 12
Poor Adequate Good Excellent

60 - 63 E ☐ 01 ☐ 02 ☐ 03 ☐ 04 ☐ 05 ☐ 06 ☐ 07 ☐ 08 ☐ 09 ☐ 10 ☐ 11 ☐ 12
Poor Adequate Good Excellent

Factor IV: Overall effectiveness in the Patient Management Conference

Candidates who rank HIGH have made an overall good impression in terms of the interaction of Factors I, and III. Candidates may rank moderate in I, II, and III, but high in IV because the interaction of all three factors produces a favorable impression.

Candidates who rank LOW have done so poorly in one or more of the other factors as to render their performance on the whole as ineffective.

A ☐ 01 ☐ 02 ☐ 03 ☐ 04 ☐ 05 ☐ 06 ☐ 07 ☐ 08 ☐ 09 ☐ 10 ☐ 11 ☐ 12
Poor Adequate Good Excellent

B ☐ 01 ☐ 02 ☐ 03 ☐ 04 ☐ 05 ☐ 06 ☐ 07 ☐ 08 ☐ 09 ☐ 10 ☐ 11 ☐ 12
Poor Adequate Good Excellent

C ☐ 01 ☐ 02 ☐ 03 ☐ 04 ☐ 05 ☐ 06 ☐ 07 ☐ 08 ☐ 09 ☐ 10 ☐ 11 ☐ 12
Poor Adequate Good Excellent

D ☐ 01 ☐ 02 ☐ 03 ☐ 04 ☐ 05 ☐ 06 ☐ 07 ☐ 08 ☐ 09 ☐ 10 ☐ 11 ☐ 12
Poor Adequate Good Excellent

E ☐ 01 ☐ 02 ☐ 03 ☐ 04 ☐ 05 ☐ 06 ☐ 07 ☐ 08 ☐ 09 ☐ 10 ☐ 11 ☐ 12
Poor Adequate Good Excellent

COMMENTS:

THE AMERICAN BOARD OF ORTHOPAEDIC SURGERY
ORAL EXAMINATION RATING FORM

COLUMNS

IDENTIFICATION

1 -14 Candidate's Identification Number ____________________

15-24 Examiner's Name ____________________ Alternate ____________________

25 Subject: 1 ☐ Adult Orthopaedics 3 ☐ Trauma

2 ☐ Children's Orthopaedics 4 ☐ Interpretive Skills

5 ☐ Simulations

26-35 Date ____________________ Starting Time ____________________

36-55 Cases ________ ________ ________ ________ ________

EVALUATION

Columns		Unable To Evaluate	Definite Failure	Marginal	Good	Excellent
56-57	Recall of Factual Information	☐ 00	☐ 01 ☐ 02 ☐ 03	☐ 04 ☐ 05 ☐ 06	☐ 07 ☐ 08 ☐ 09	☐ 10 ☐ 11 ☐ 12
58-59	Analysis and Interpretation of Clinical Data	☐ 00	☐ 01 ☐ 02 ☐ 03	☐ 04 ☐ 05 ☐ 06	☐ 07 ☐ 08 ☐ 09	☐ 10 ☐ 11 ☐ 12
60-61	Problem-Solving Ability; Clinical Judgment	☐ 00	☐ 01 ☐ 02 ☐ 03	☐ 04 ☐ 05 ☐ 06	☐ 07 ☐ 08 ☐ 09	☐ 10 ☐ 11 ☐ 12
62-63	Relates Effectively; Shows Desirable Attitudes	☐ 00	☐ 01 ☐ 02 ☐ 03	☐ 04 ☐ 05 ☐ 06	☐ 07 ☐ 08 ☐ 09	☐ 10 ☐ 11 ☐ 12

COMMENTS

64-69 The Candidate was difficult to evaluate because:

☐ 64 He spoke slowly

☐ 66 He spoke rapidly

☐ 68 He did not speak English well

☐ 65 He seems excessively nervous

☐ 61 He seemed confused about the examination procedure

☐ 69 Other ____________________

____

70 ☐ I did not find the Candidate difficult to evaluate

71 REMARKS:

Appendix 14

RESULTS OF MULTIPLE CORRELATIONAL ANALYSIS USING SUBSCORES AS INDEPENDENT VARIABLES

NOTE: In reading this table please note that the following abbreviations are to identify the tests, sub-tests and listed as independent variables.

Test or Sub-test

DI = Diagnostic Interview
MC = Multiple Choice
Prob = Problem Identification
PTI = Proposed Treatment Interview (oral)
WS = Written Simulation

Bio Mech = Biomechanics
Gen Orth = General Orthopedics
Hand Surg = Hand Surgery

Score

Diag. Sel. = Selection of Indicated Procedures on Diagnosti Problems
Diag. Avoid = Avoidance of Contra-Indicated Procedures o Diagnostic Problems
Treat. Sel. = Selection of Indicated Procedures on Treatment Problems
Treat. Avoid = Avoidance of Contra-Indicated Procedures o Treatment Problems

Appendix 14 (con't)

RESULTS OF MULTIPLE CORRELATIONAL ANALYSIS USING SUBSCORES AS INDEPENDENT VARIABLES

First and Second Year Residents
N = 109

Dependent Variable	R	F	Independent Variables	Partial r	F	Simple Correlation
Factual Information	.53	1.83*				
			Prob I WS Diag Avoid	-.28	7.48	**-.19
			PTI Overall	.28	7.42	** .12
			PTI Interaction	-.26	6.46	* .04
			Prob II WS Treat Avoid	.24	5.20	* .08
			Prob I WS Diag Sel	-.22	4.71	* -.10
			Prob II WS Total Sel	-.18	3.04	-.06
Problem Solving	.51	1.67				
			Prob II WS Diag Avoid	-.27	7.02	**-.24
			PTI Overall	.28	5.56	* .14
			PTI Manner	-.24	5.24	* .01
			MC Trauma	.20	3.71	.23
			MC Bio Mech	-.20	3.61	-.06
			Prob I Diag Sel	-.16	2.28	-.02
Information Gathering	.42	1.02				
			Prob II WS Diag Avoid	-.22	4.46	* -.17
			PTI Overall	.20	3.63	.14
			PTI Manner	-.20	2.42	.04
			MC Bio Mech	-.16	2.35	-.05

APPENDIX 14 (con't)

RESULTS OF MULTIPLE CORRELATIONAL ANALYSIS USING SUBSCORES AS INDEPENDENT VARIABLES

First and Second Year Residents
N = 109

Dependent Variables	R	F	Independent Variables	Partial r	F		Simple r
Clinical Judgment	.45	1.19					
			PTI Overall	.25	5.92	*	.09
			Prob II WS Diag Avoid	-.18	3.12		-.16
			PTI Interaction	-.18	3.05		.03
			PTI Manner	-.17	2.50		-.01
			Prob I WS Diag Sel	-.16	2.25		-.05
Patient Relations	.45	1.21					
			Prob II WS Treat Sel	.21	4.10	*	.18
			MC Bio Mech	-.21	3.79		-.08
			MC Trauma	.19	3.40		.15
			Prob II Diag Avoid	-.18	3.01		.17
			PTI Manner	-.16	2.40		-.01
			Prob I WS Diag Sel	-.16	2.39		-.04
Colleague Relations	.51	1.62					
			Prob II WS Diag Avoid	-.25	5.98	*	-.21
			PTI Overall	.23	4.95	*	.19
			MC Bio Mech	-.21	4.11	*	-.08
			Prob I Diag Sel	-.19	3.35		-.02
			MC Trauma	.18	3.06		.17
			PTI Manner	-.17	2.78		.07
			Prob I WS Diag Avoid	-.16	2.32		-.11

APPENDIX 14 (con't)

RESULTS OF MULTIPLE CORRELATIONAL ANALYSIS USING SUBSCORES AS INDEPENDENT VARIABLES

First and Second Year Residents
N = 109

Dependent Variable	R	F	Independent Variables	Partial r	F	Simple r
Ethics	.55	2.01	*			
			PTI Manner	-.31	9.14	**-.08
			Prob II Treat Sel	.28	7.28	** .19
			Prob I Diag Sel	-.21	4.09	* -.05
			MC Trauma	.21	3.85	.21
			MC Bio Mech	-.19	3.30	-.09
Overall	.48	1.37				
			PTI Overall	.27	6.95	** .12
			Prob II Diag Avoid	-.27	6.79	**-.18
			Prob I Diag Avoid	-.21	4.13	* .11
			PTI Manner	-.19	3.24	.01
			MC Gen Orth	.17	2.77	.18
			DI Interaction	-.17	2.66	.06

APPENDIX 14 (con't)

RESULTS OF MULTIPLE CORRELATIONAL ANALYSIS USING SUBSCORES AS INDEPENDENT VARIABLES

Third and Fourth Year Residents
N = 119

Dependent Variables	R	F	Independent Variables	Partial r	F	Simple r
Factual Information	.53	1.60				
			MC Gen Orth	-.20	4.11	* -.20
			MC Bio Mech	.20	4.06	* .20
			Adult Oral	.17	2.96	.17
			MC Pathology	.17	2.92	.17
			Prob I Treat Avoid	-.17	2.67	-.17
			MC Hand Surg	.16	2.52	.16
Problem Solving	.56	1.85	*			
			Adult Oral	.22	4.87	* .35
			Prob I WS Diag Avoid	.20	4.02	* .30
			Prob I WS Treat Avoid	-.19	3.37	-.09
Information Gathering	.63	2.70	**			
			DI Diagnosis	.21	4.41	*
			Prob I WS Treat Avoid	-.20	3.85	
			Adult Oral	.20	3.74	
			MC Bio Mech	.16	2.53	
			Prob I WS Diag Avoid	.15	2.27	
			MC Hand Surg	.15	2.22	

APPENDIX 14 (con't)

RESULTS OF MULTIPLE CORRELATIONAL ANALYSIS USING SUBSCORES AS INDEPENDENT VARIABLES

Third and Fourth Year Residents
N = 119

Dependent Variables	R	F	Independent Variables	Partial r	F	Simple r
Patient and Relationships	.49	1.29				
			Adult Oral	.24	5.84	* .23
			Prob II WS Treat Avoid	-.19	3.59	-.10
			DI Communication	-.18	3.15	-.12
			DI Diagnosis	.16	2.50	.05
			MC Hand Surgery	.15	2.07	.13
Colleague Relationships	.53	1.58				
			Prob I WS Treat Avoid	-.23	5.06	* -.13
			DI Communication	-.17	2.80	-.02
Overall Competence	.41	1.48				
			MC Bio Mech	.19	3.66	.33
			Adult Oral	.17	2.73	.27
			MC Gen Orth	-.16	2.37	.16
			MC Hand Surg	.15	2.09	.21

Note: Independent Variables with F-ratios below 2.00 are not shown even though they make some contribution to the R. Some dependent variables have not been shown.

* Sig at .05 level
** Sig at .01 level
PTI = Proposed Treatment Interview

Appendix 15

Multiple Regression Analysis, 1968 Certifying Examination

NOTE: In reading this table note that the following abbreviations are used to identify the tests, sub-tests and score listed as independent variables. In each case the variable name lists first the content or form of the test with a dash followed by the name of the score or subscore. All tests other than those specifically identified as multiple choice or written simulations are in the form of Oral exercises.

Tests or sub-tests

MC = Multiple Choice
O and I = Observation and Interpretive Skills
WS = Written Simulation

Scores

(same as 14)
+
PS = Problem Solving

APPENDIX 15 (con't)

N = 391

Dependent Variable	Reliability of Dependent Variable	R	F	Independent Variables	Partial r	F	Simple r
Information Gathering	(.29)	.36	5.13**	Multiple Choice-Recall	.12	4.48*	.21
				O and I-Interpretation	.11	4.48*	.21
				Trauma-Problem Solving	.10	3.84	.22
				Child-Problem Solving	.08	2.43	.17
				MC-Problem Solving	.06	1.34	.19
				Adult-Problem Solving	.05	0.82	.16
				Simulation-Attitudes	.05	0.82	.14
				WS-Treat Select	.04	0.81	.12
				WS-Diagnostic Select	.04	0.48	.15
				WS-Treat Avoid	-.03	0.23	.06
				WS-Diagnostic Avoid	.02	0.12	-.14
Problem Solving	(.29)	.40	6.72**	O and I-Interpretation	.16	9.81**	.27
				Multiple Choice-Recall	.14	7.56**	.28
				Simulation-Attitude	.09	3.31	.19
				Trauma-Problem Solving	.08	2.48	.21
				MC-Problem Solving	.08	2.47	.23
				WS-Treat Select	.07	2.03	.14
				Adult-Problem Solving	.06	1.39	.18
				Child-Problem Solving	.03	0.32	.14
				WS-Treat Avoid	-.12	0.10	.07
Clinical Judgment	(.22)	.34	4.46**	O and I-Interpretation	.13	6.21*	.22
				Trauma-Problem Solving	.09	2.79	.20
				Multiple Choice-Recall	.08	2.62	.21
				Simulation-Attitudes	.08	2.46	.17
				Adult-Problem Solving	.06	1.24	.16
				WS-Treat Select	.05	1.01	.13
				WS-Diagnostic Select	.04	0.69	.15
				WS-Problem Solving	.04	0.55	.16
				WS-Diagnostic Avoid	.01	.03	-.04
				WS-Treat Avoid	-.01	.02	.06

APPENDIX 15 (con't)

N=391

Dependent Variable	RELIABILITY TO DEPENDENT VARIABL	R	F	INDEPENDENT VARIABLES	PARTIAL R	F	SIMPLE
Surgical Technique	(.29)	.26	2.42**	O and I-Interpretation	.12	5.39*	.18
				Child-Problem Solving	.08	2.28	.14
				Adult-Problem Solving	.07	1.94	.14
				MC-Problem Solving	.04	0.59	.12
				Trauma-Problem Solving	.04	0.51	.12
				WS-Treat Select	.04	0.48	.08
				MC-Recall	.04	0.46	.13
				Simulation-Attitudes	.03	0.30	.10
				WS-Diagnostic Avoid	.01	0.84	-.01
				WS-Diagnostic Select	-.01	0.03	.06
Patient Relations	(.25)	.27	2.80**	Simulation-Attitudes	.11	4.35*	.15
				WS-Diagnostic Select	.11	4.20*	.18
				O and I-Interpretation	.10	3.78	.17
				Adult-Problem Solving	.06	1.30	.12
				WS-Treat Select	.05	0.88	.14
				WS-Diagnostic Avoid	.03	0.39	-.05
				WS-Treat Avoid	-.02	0.15	.02
				Trauma-Problem Solving	.02	0.12	.10
				MC-Problem Solving	.01	0.05	.09
				MC-Recall	.01	0.03	.10
				Child-Problem Solving	.01	0.03	.09
Continuing Responsibility	(.23)	.29	3.20**	O and I-Interpretation	.10	3.54	.18
				WS-Diagnostic Select	.10	3.42	.17
				Adult-Problem Solving	.07	2.10	.15
				Child-Problem Solving	.07	1.93	.15
				Simulation-Attitudes	.05	1.08	.12
				MC-Recall	.05	.01	.16
				WS-Treat Avoid	.05	.077	.09
				WS-Treat Select	.04	0.49	.13
				WS-Diagnostic Avoid	.04	0.49	.1[illegible]
				WS-Diagnostic Avoid	.03	0.30	-.03
				MC-Problem Solving	.01	0.05	.11

APPENDIX 15 (cont'd)

N = 391

Dependent Variable	Reliability of Dependent Variable	R	F	Independent Variables	Partial r	F	Simple r
Emergency Care	(.28)	.28	3.0**	O and I-Interpretation	.14	7.85**	.21
				Trauma-Problem Solving	.07	1.67	.16
				Adult-Problem Solving	.06	1.42	.14
				WS-Diagnostic Select	.06	1.38	.11
				Child-Problem Solving	.06	1.25	.13
				WS-Diagnostic Avoid	.04	0.63	-.02
				MC-Problem Solving	.04	0.58	.14
				Simulation-Attitudes	.04	0.55	.12
				Multiple Choice-Recall	.04	0.53	.15
				WS-Treat Select	-.03	0.42	.04
				WS-Treat Avoid	-.03	0.40	.03
Colleague Relations	(.26)	.26	2.5**	O and I-Interpretation	.07	3.31	.16
				WS-Treat Select	.08	2.50	.14
				Adult-Problem Solving	.07	2.23	.14
				Simulation-Attitudes	.07	1.78	.12
				Multiple Choice-Recall	.06	1.44	.14
				WS-Treat Avoid	-.06	1.31	-.01
				WS-Diagnostic Avoid	.04	0.66	-.01
				WS-Diagnostic Select	.03	0.45	.12
				MC-Problem Solving	.03	0.33	.12
				Trauma-Problem Solving	.02	0.15	.10
Ethics	(.28)	.25	2.41**	Adult-Problem Solving	.09	3.05	.16
				Multiple Choice-Recall	.07	1.60	.16
				WS-Treat Select	.06	1.39	.11
				Simulation-Attitudes	.06	1.19	.12
				Child-Problem Solving	.06	0.89	.12
				Trauma-Problem Solving	.05	0.73	.14
				MC-Problem Solving	.04	.60	.13
				WS-Treat Avoid	-.04	.47	.02
				WS-Diagnostic Select	.03	.35	.12
				O and I-Interpretation	.03	.26	.11
				WS-Diagnostic Avoid	.02	.09	-.04

APPENDIX 15 (cont'd)

N = 391

Dependent Variable	Reliability of Dependent Variable	R	F	Independent Variables	Partial r	F	Simple r
Overall Competence	.31	.37	5.28**	O and I-Interpretation	.12	5.48*	.22
				Multiple Choice-Recall	.12	5.46*	.26
				Multiple Choice-PS	.09	3.13	.22
				Adult-Problem Solving	.08	2.24	.18
				Simulation-Attitudes	.06	1.21	.15
				Trauma-Problem Solving	.05	.86	.18
				WS-Treat Select	.05	.84	.12
				Child-Problem Solving	.05	.77	.14
				WS-Diagnostic Select	.03	.70	.15
				WS-Diagnostic Avoid	.02	.19	-.03
				WS-Treat Avoid	-.01	.02	.07

* Sig at .05 level

** Sig at .01 level

Appendix 16

American Board of Orthopaedic Surgery

INSTRUCTIONS TO EXAMINERS FOR ORAL EXAMINATIONS

INTRODUCTION

The oral examinations will consist of 5 one-half hour examinations. Three of these--Problem-Solving Adult, Problem-Solving Children's and Problem-Solving Trauma--will focus on the candidate's ability to handle realistic clinical problems. The fourth, the Simulated Interview, will focus on ability to relate effectively to patients and colleagues. The fifth, Observation and Interpretation, will focus on ability to observe and interpret data. Detailed instructions for administering and evaluating each of these examinations are given in the following sections of this document. A separate document entitled Instructions to Candidates is enclosed so that you will have an opportunity to review the information the candidates have been given about the examinations.

This document describes all of the oral examinations. However, you will wish to give special attention to the sections relating to the examination you are administering and the general instructions on rating. Copies of all the cases you will be using will be made available the night before the oral examinations.

Prepared with the Assistance of

The Center for the Study of Medical Education
University of Illinois College of Medicine

PROBLEM-SOLVING ADULT, CHILDREN'S AND TRAUMA

The Process of Problem Solving

The main purpose of these examinations is to evaluate the candidate's ability to reason correctly and logically and to arrive at a diagnosis or plan of treatment for a particular patient based upon a consideration of all criteria.

The analysis of the oral examinations given in the past revealed that they often duplicated the written examination in testing for knowledge content (recall, remembering). This important area can be tested by written examinations, and it would seem that the hundreds of man hours invested in orals could be used more profitably to obtain information about the candidates thought processes, skills and attitudes that cannot be readily assessed in the written format.

Since the majority of a candidate's knowledge is intended for application to problems in real life, our task in evaluating his problem-solving ability will be easier if the examiner starts with the premise that possession of knowledge and the ability to apply it are not synonymous. It may also help to consider the steps in problem solving:

(1) The problem is identified. (History, examination, laboratory data).

(2) Depending upon his familiarity with such problems, the candidate perceives the problem as:

   A. Immediately having familiar aspects to guide thinking.

   B. Initially unfamiliar so he searches for familiar elements.

(3) Reconstructs familiar elements to make them more completely resemble a familiar patient problem.

(4) Reinterpretation of the patient problem in light of ALL available data about the patient and the present state of knowledge. (clinical judgement)

(5) Selects "orthopaedic principle," theory, idea or method of generalization suitable to problem.

(6) Applies principle to problem. (tentative diagnosis or treatment)

(7) Arrives at solution and confirms it. (working diagnosis or treatment)

While this process is based upon the candidate's possession of basic knowledge, in this examination we are interested in how he applies whatever knowledge he has. If it is a diagnostic problem, does he "jump" to conclusions or does he systematically rule out alternative possibilities. In the therapeutic problems does he have the ability to reason effectively in supporting his decisions about treating a case for which there is not necessarily one acceptable solution.

Formats to be Used

The examinations in Adult and Children's will include two types of problems. The Diagnostic Problem will require the candidate to elicit information concerning a particular patient from the examiner and then present his conclusions and the reasons supporting them. The second type, the Defense of Therapy Problem, requires the candidate to review the findings concerning a patient and outline and defend a course of treatment.

In case a candidate either "blocks" on a problem or solves it with extraordinary dispatch, the examiner will have one additional Defense of Therapy Problem available.

The Trauma examination will include the Emergency Treatment and Complication Problems. The Emergency Treatment Problem will require the candidate to outline his treatment of an emergency patient with multiple injuries. The Complication Problem will require the candidate to describe and defend his management method. The Trauma examiners will have an additional Complication Problem to use in case a candidate "blocks" on a problem or finishes his problems very quickly.

Procedures to be Followed

Diagnostic Problem

At the beginning of the examination you should hand the candidate the case description which indicates the age, general appearance, occupation and chief complaint of a patient. You should instruct the candidate that his task is to elicit data on the history, physical examination, laboratory findings and x-ray findings from you.

Unlike some earlier experiments with this type of exercise, you will not be "role playing" a patient during the inquiries on the history, but you will simply give the historical findings as requested. IT IS EXTREMELY IMPORTANT, however, that you insist that the candidate be specific in his inquiries, and you refuse to answer any vague questions about the "general condition" of the patient. Be very careful about giving irrelevant clues to the candidate. If, for example, he asks, "Does the patient have an earlier injury to his elbow?", do not say, "Yes." But say, "The patient says he hurt his arm when he was seven years old." It will then be up to the candidate to find out if this arm injury was of the elbow, or some other fracture.

It is recognized that the candidates will from time to time ask unanticipated questions. Answers to these questions will have to be "ad libbed" on the spot so as to fit as precisely as possible the case described. If the candidate asks pertinent questions for which you have not supplied an answer, give an answer consistent with the case and diagnosis. Part of the test is, after all, designed to determine the candidate's ability to elicit the proper information regarding the case. If, on the other hand, the unanticipated questions are irrelevant or immaterial, you will have to answer in a vague or non-specific way. Sometimes a simple "I don't know" is best.

You will be supplied x-rays. If the candidate requests an x-ray which is not available, you may describe any abnormalities that would have been present.

After about 10 minutes of information gathering you should stop the candidate and ask him for his diagnostic impressions and his reasons for preferring this diagnosis over other possibilities. You may then ask a few more questions designed to probe the candidate's mental processes. You should not, however, engage in a"debate" with the candidate as more information is probably obtainable by exposing him to another problem.

Defense of Therapy Problem

At the beginning of the examination you should give the candidate the description of the case. The description will inform him that he may obtain additional pertinent information from you. You should instruct him that his task is to formulate a definite treatment plan for the patient and to explain his reasons for recommending it. You should allow him about 3 minutes to read the description and about 10 minutes to describe his procedures.

You should question his recommendations and conclusions in order to discover the criteria the candidate is using to arrive at the solution and the rationale upon which his recommendations are based. Remember, you are not so much interested in what method of treatment he uses as in his reasons for choosing that method. Does he use available data and orthopaedic principles, or is he a "cookbook orthopaedist" who has one answer for a given situation and doesn't want to be "confused by the facts"?

In discussing the case with the candidate you should strive to put him at ease by avoiding a threatening manner and emotional dialogue. You should avoid giving him clues as to what you think is the optimal course.

You can control the tempo of the examination by keeping in mind the following criteria for judging ANY treatment the candidate elects.

Has he identified the problem?
Does he understand it in his own terms? (categorized into the proper "model system")
Did he have the proper data?
Did he ask for more pertinent data?
Does he use available knowledge about the condition?
What does he expect to accomplish?
Is this based upon the patient's needs?
Does he select the proper principles to apply to this specific problem?
Can he distinguish between scientifically discovered and proven principles and empirically based ones?
Does he know which type he is using?
Did he weight various factors properly?
What were his reasons for weighting factors low or high in arriving at a therapeutic plan?
Could he apply his method to this given case?
Will it accomplish the desired result?
Will it be the method of least risk to the patient?
Does he anticipate complications?
How will he deal with them?
Is he aware of alternate plans?
Why not use them?
What if his plan fails?
How will he judge the end result (criteria)?

Your questions can then be designed to see how well candidates meet these criteria.

Emergency Treatment Problem

The main purpose of this examination is to gain information on the candidate's understanding of the most effective ways in which he can meet his responsibilities as a physician and an orthopaedist in providing emergency care for multiple injuries.

You should give the candidate the case description and instruct him to outline a diagnostic and treatment program for the patient described, and to indicate what priorities he would establish for each step in his plan. You should permit him to outline his entire plan of evaluation and treatment up to such time as non-orthopaedic consultation is available. If he attempts to solve the problem by resorting to immediate consultation, suggest to him that no consultation is immediately available.

After the candidate has outlined his plan you may then attempt to ascertain his reasons for these recommendations and the priorities he would assign to each component. You may give him feedback as to the results of various moves as he goes along so that he can use such data in deciding on his subsequent moves. You should allow the candidate about 3 minutes to read the case description, and about 10 minutes for the discussion.

You should avoid giving the candidate clues by pointing out errors in priority or action; instead, focus on the reasoning behind his recommendations. You are attempting to find out WHY he chooses a given step at a particular time in the management of the case. Does he establish priority of treatment based upon knowledge, data at hand and judgement as to the needs of this particular patient at this time, or is he simply following a set routine?

This technique in many ways is a variant of the Defense of Therapy Problem. The instructions for that technique also apply to this one.

Complication Problem

The main purpose of this examination is to obtain information on the candidate's ability to formulate a plan of management for a patient in which some complications have developed. The candidate need not be expert in the detailed management of such illness, but the successful candidate should be able to indicate the most effective methods of defining the patient's problem and organizing effective help into a plan of treatment.

At the beginning of the examination give the candidate the description of the problem, and inform hin that his task is to outline a plan of management including further diagnostic and therapeutic measures he believes necessary. You should provide him with feedback on the results of various steps in his management. You should allow the candidate about 3 minutes to study the case description, and about 10 minutes to discuss his procedures.

This technique, like the *Emergency Treatment Problem*, is variant to the *Defense of Therapy Problem*. The instructions regarding that technique also apply to this one.

Administration of the Examinations

Adult and Children's Examination

The first problem used should be the *Diagnostic Problem*. When the candidate enters the room, check his identification and give him the case description of the *Diagnostic Problem*. Allow him about a minute to review the sheet and then ask him to proceed. Give him about 10 minutes and then ask him for his diagnostic impressions. If he wishes to give you his impressions earlier, he may do so. He should be given about 3 minutes to describe his impressions and to answer any questions you may pose. You will then give him the the Candidate's Case Description of the *Defense of Therapy Problem* and allow him about 3 minutes to read it. He should then proceed to describe his suggestions for therapy. These suggestions should take about 10 minutes. At this point stop the examination and dismiss him. After he leaves, mark the Rating Form and write any pertinent notes or comments on the back. If the candidate "blocks" on the *Defense of Therapy Problem* or finishes both problems very quickly, use the back-up *Defense of Therapy Problem* to fill the remaining time.

Trauma Examination

The first problem used will be the *Emergency Treatment Problem*. When the candidate enters the room, check his identification, give him the *Case Description Emergency Treatment Problem* and allow him about 3 minutes to read the problem. He should then proceed with his description of the therapeutic steps he would follow. These suggestions should take about 10 minutes. At the end of this time, you should hand him the Case Description of the *Complication Problem* and give him about 3 minutes to read it. He should then be given about 10 minutes more to discuss the problem with you. At this point stop the examination and dismiss him. After he leaves, mark the Rating Form and write any pertinent notes or comments on the back. If the candidate "blocks" on any problem or finishes both problems very quickly, use the back-up *Complication Problem* to fill the remaining time.

Rating the Problem Solving Examinations

General

All of the candidates in all examination subjects are to be rated on all four factors described in the special rating forms supplied by the Board. However, some factors will be much more easily rated than others in this examination. The Examination Committee of the Board will take this into account in deciding on the appropriate weights to be assigned for each factor in the examination. If you believe that you simply cannot rate a candidate on a particular factor in an examination, check the box entitled "Unable to rate" on the Rating Form. Note that the main emphasis in the Problem-Solving Adult, Children's and Trauma portions of the examination is on problem solving and clinical judgement, as described on the Rating Form and in the special notes below. You should conduct your examinations to arrive at a clear impression of the candidate's rating on this factor. You should not be too greatly concerned about other factors; they are being extensively probed in other parts of the examination. Do not, however, use the "unable to rate" option unless you have absolutely no impression of the candidate's ability in these areas. If, for example, he fails to read an x-ray properly, mark him as a failure in factor 2, "ability to analyze and interpret clinical data." Your impressions which by themselves are unreliable, can, when combined with other data, serve to give a reliable overall picture of the candidate's ability.

In addition to the notes below, be sure to carefully review the Rating Form and the General Comments on Rating.

Notes on Rating Problem Solving for Each Technique Used.

Diagnostic Problem. In rating this examination the examiner should keep in mind that reasonable, efficient thoroughness in gathering date, coupled with intelligent use of the data so gathered in arriving at a realistic primary diagnosis, are the keys to the candidate's success in this part of the examination. The cases have been selected to avoid both obvious, straighforward problems and rare, unusual cases. The candidate's approach is extremely important. Not all of us are so mechanically efficient that we don't make occasional false starts or take an occasional wrong turn in working out a diagnostic problem. However, we must realize that gross inefficiency can waste so much time as to lead to decreases in the quality of patient care that can be delivered by the medical profession. Furthermore, the ability to acquire data loses much of its value if the information acquired is not synthesized into some realistic and meaningful conclusion. However, the emphasis in this portion of the examination should be as much on the candidate's methods of arriving at a diagnosis as it is on his obtaining a precisely correct diagnosis.

Defense of Therapy Problem. In this examination the most acceptable candidate will be able to:

(1) recognize the basic problem,
(2) make requests for further data which are reasonable in light of the basic problem,
(3) base his reasoning on an accurate comprehension of relevant facts and generally accepted principles,
(4) choose realistic, effective and practical therapeutic solutions in the light of the data presented,
(5) demonstrate ability to respond flexibly and reasonably to questions about the handling of complications of failure of initial therapy,
(6) choose valid critera to judge results.

The unacceptable candidate will:

(1) be unable to recognize the basic problem,
(2) request additional data in a "shotgun" manner,
(3) treat the case on the basis of indications also, not recognizing the need for prerequisite conditions,
(4) suggest treatment regimens tailored to his own preconcieved notions rather than to a particular patient,
(5) adhere to a rigid or ineffective approach,
(6) fail to judge results well.

Emergency Treatment Problem. The acceptable candidate in this technique should be able to outline a realistic program to evaluate adequately any potentially lifethreatening situations, to manage such problems initially in a way that minimizes dangers to the patient, and to initiate realistic treatment or diagnostic routines that will lay the foundations for later management of other problems. The inadequate candidate will not recognize the problem or will delay evaluation and management of the problem to the point of jeopardizing the patient's life.

In explaining his reasons for the treatment selected, the adequate candidate will have sound reasons for his procedures in contrast to the "treatment by rote" exhibited by the individual who simply follows a prescribed routine.

The adequate candidate will also reveal his problem-solving skills by asking for relevant information about the patient's condition and by using this information in the course of his discussion of dianosis and treatment. He will be alert to possible complications and to occult injuries of significance. In contrast, the inadequate candidate will be oblivious to the implications of data presented, may waste time or complicate matters by recommending time-consuming diagnostic procedures, or procedures that require excessive manipulation of the acutely injured patient.

Complication Problems. In rating this technique, the acceptable candidate will:

(1) correctly appraise the problem,
(2) choose a realistic plan in the light of data presented,
(3) continue to follow the patient,
(4) be aware of possible complications.

The unacceptable candidate will:

(1) not recognize the problem,
(2) treat the case in an unreasoning conditioned manner,
(3) fail to follow the patient,
(4) fail to recognize complications.

SIMULATED INTERVIEWS

Description of the Examination

The critical incident study of the critical performance requirements for orthopaedic surgeons listed a number of requirements which dealt with the orthopaedist's ability to handle situations requiring interaction between himself and patients and himself and colleagues. It has been found through extensive research and experimentation that the most effective and efficient way to gather such information is through role playing in which you play the role of a patient, a consulting physician or nurse, or other member of the health team, and the candidate plays the role of an orthopaedic surgeon.

In each situation the candidate will be given a description of a situation based upon a clinical case history. He will then be expected to play the role of a physician in two or three typical situations such as:

(1) explaining the next step in diagnosis or treatment,
(2) discharging a patient from the hospital,
(3) explaining to a patient that he has found no abnormality,
(4) discussing a poor prognosis with the patient or family,
(5) talking to a patient who has consulted other orthopaedists,
(6) talking to a nurse about a change in procedures.

Some Suggestions on Role Playing

In this interview it is particularly important that the examiner act as a typical person of the age, sex, educational and occupational level described.

In order to assist in keeping the interview moving and to assure that each candidate faces a comparable situation, some suggested questions are noted below. This list is not exhaustive or appropriate for every situation. Use it as a guideline. The examination will be most effective if you place yourself in the situation and ask the questions the candidate's statements would naturally evoke at the point in the discussion when each seems most appropriate.

Questions that would often be asked by the typical patient:

1. What is wrong with me?
2. Why do I have to go to the hospital?
3. Can I return to my usual occupation?
4. If I follow the average course what can I expect?
5. How long will I be off work?
6. Do I need an operation?
7. What is going to be done at the hospital?
8. Can I have another opinion?
9. Do you think I should go to a chiropractor?
10. May this recur?
11. Might this be a cancer?
12. Is this treatment dangerous?
13. Do I have arthritis?

Questions that would often be asked by a mother regarding treatment of her child:

1. Will he be able to walk?
2. Will he have any deformity?
3. Are all of these x-rays necessary?
4. Does this run in the family?
5. Why can't I stay with the child?
6. Can we wait a while?
7. Will he be normal?

Administering the Simulated Interviews.

This examination will be administered by two examiners who will use three simulations during each half hour. When the candidate enters the room he should be introduced to both examiners and then given the case description of the first simulation. He should be given approximately 3 minutes to read the case description and then he should start the simulation. At the end of approximately 6 minutes the observing examiner should call time and hand the candidate a new simulation. At this point, the examiner who observed the last simulation will "role play" the next one. This simulation should be administered in the same fashion as the first one. When the second simulation is completed the examiners will again change places and administer a third simulation.

At the end of the third simulation the bell will ring and the candidate should proceed to the next examination. At this point both examiners should mark their Rating Forms independently.

Rating the Simulated Interviews

General

All of the candidates in all examination subjects are to be rated on all four factors described on the special rating forms supplied by the Board. However, some factors will be much more easily rated than others in this examination. The Examination Committee of the Board will take this into account in deciding on the appropriate weights to be assigned to each factor in this examination. If you believe that you simply cannot rate a candidate on a particular factor, check the box entitled "Unable to Rate" on the Rating Form. Note that the main emphasis in the Simulated Interview is on ability to relate effectively, as described on the Rating Form and in the special note below. Conduct your examination to arrive at a clear impression of the candidate's rating on this factor. Do not be too greatly concerned about the other factors; they are being extensively probed in other parts of the examination. Do not, however, use the "Unable to Rate" option unless you have absolutely no impression of the candidate's ability in these areas. If, for example, he fails to read an x-ray properly, mark him as a failure in factor 2, "ability to analyze and interpret clinical data." Your impressions, which by themselves are unreliable, can, when combined with other data, serve to give a reliable overall picture of the candidate's ability.

In addition to the notes below, be sure to carefully review the Rating Form and the General Comments on Rating.

Note on Rating Ability to Relate in the Simulated Interviews

This factor contains a number of elements which you should keep in mind in evaluating the candidate's performance. First, you should consider the information the candidate provides in terms of its effectiveness in meeting the goals of the interview. Candidates can do poorly in a number of ways.

(1) They can say things which would cause needless alarm, discomfort or embarrassment. For example, they can over-emphasize the consequences of lack of treatment.

(2) They can fail to speak honestly to the patient. For example, they can make over-optimistic claims regarding the effectiveness of treatment.

(3) They can make statements which reveal unprofessional attitudes toward patients, colleagues or other medical personnel, or could be construed as tactless, indiscreet or undignified.

Second, you should consider the manner in which the candidate conducts himself in the interview. Does the way he conducts the interview, in terms of posture, voice, mannerisms, illustrations, gestures, etc., communicate genuine concern and interest in the problems of the person involved.

OBSERVATION AND INTERPRETATION

General Description of the Examination

The ability to observe and interpret accurately is an integral part of the requirements for the successful practice of orthopaedic surgery. The conclusions drawn from direct observation frequently determine the diagnosis, decide the course of treatment or determine the efficacy of treatment. This examination is directed toward the evaluation of the candidate's ability to observe and correlate information derived from microscopic slides and x-rays.

The examination will last approximately one-half hour and will consist of 5 or 6 sets of materials, some of which will involve the interpretation of x-rays, some will deal with the interpretation of pathology slides, and others will require the correlation of slides with x-rays. For each set of materials the emphasis will be on the candidate's accuracy of observation and interpretation of what he sees.

Administration of the Observation and Interpretation Examination

You should first present the material to the candidate and instruct him to describe what he sees precisely as he might in a written report, indicating any abnormalities that may be present.

If the candidate fails to interpret the material properly you might supply him with some additional historical, physical examination or laboratory data which would assist him. You should not provide this additional information until AFTER he has initially described his findings. If he identifies some abnormalities you should then ask additional questions which would probe his ability to interpret what he sees in the light of his knowledge and understanding of physiological and pathological processes. For example, you might ask him to speculate as to the reason that the structures on the slide show the patterns they do, or you might ask the probable effect on the abnormality of various types of treatment.

DO NOT SPEND TOO MUCH TIME ON ANY ONE EXERCISE. Remember, you are mainly concerned with what the candidate sees and how he interprets it. If you ask too many questions about diagnosis the candidate may simple answer them on the basis of his basic information and not on the basis of any observational skills. Although we recognize that it is extremely important to assess the candidate's basic store of information on diagnosis and treatment, this area of competence is being probed thoroughly in other portions of the examination.

Rating the Observation and Interpretation Examination

All of the candidates in all examination subjects are to be rated on all of the four factors described on the special rating forms supplied by the Board. However, some factors will be much more easily rated than others in this examination. The Examination Committee of the Board will take this into account in deciding on the appropriate weights to be assigned to each factor. If you believe that you simply cannot rate a candidate on a factor in a particular examination, check the box entitled "Unable to Rate" on the Rating Form. Note that the main emphasis in the Observation and Interpretation Examination is on factor 2, the ability to analyze and interpret data. Conduct your examination to arrive at a clear impression of his ability in this factor and do not be too greatly concerned about the others. They are being extensively probed in other parts of the examination. Do not, however, use the "Unable to Rate" option unless you have absolutely no impression of the candidate's ability in these areas. Your impressions, which by themselves are unreliable, can, when combined with other data, serve to give a reliable overall picture of the candidate's ability.

In addition to the above, please carefully review the Rating Form and the General Comments on Rating.

GENERAL COMMENTS ON RATING

Arriving at a Rating for the Entire Half-Hour

It is recognized that some candidates may do better on one exercise than another. Your task is to reconcile the ratings on each exercise to arrive at an overall judgement. In most cases it is probably best simply to average the ratings, but this need not always be true. It may turn out that the candidate's performance on one exercise demonstrated such effective or such poor performance that you wish to give it more weight than a mere averaging would allow. It is perfectly all right for you to do so.

The Effect of Your Rating on the Candidate's Certification

Your ratings by themselves will not serve to pass or fail the candidates. All the ratings of the oral examiners will be gathered together with the scores from the written examinations to provide the Board with a "profile" of performance of each candidate. The Board will then decide who fails, based upon the data presented. If this system is to work it is extremely important that you report your impressions of the candidate as you see him during the one half-hour. In many cases examiners tend to be lenient because they realize that in the absence of other data, failure in a half-hour test is hardly indicative of inadequacy. However, in this case your "failing" grade will not be crucial unless others find similar evidence of inadequacy. If this system is to work, *you must rate the candidate as failing if he performs inadequately on the problem given him during the half-hour you see him.*

Errors to be Avoided in Rating

The Error of Leniency or Stringency. Most raters tend to rate individuals near the upper half of the scale. If you find that most of your ratings are in the *good* and *excellent* ranges, you should review your judgements to make certain that you are not overlooking some weaknesses. Remember that a rating on any one factor is only a small part of the evaluation of each candidate. On the other hand, some raters tend to rate individuals at the bottom half of the scale. If you find that most of your scores are in the *poor* and *marginal* ranges, perhaps you should re-examine your judgements. Remember that few can do a task perfectly, but many can do tasks in a reasonably competent fashion.

The Error of Central Tendency. Making judgements is a difficult task Some judges tend to avoid the task by rating everyone average. If most of your marks are in the *marginal* and *good* categories, perhaps you are failing to take into account some individual patterns of strengths and weaknesses.

The Halo Effect. Since the factors described are interrelated, some judges tend to rate a man at the same level on all factors. Experience indicates, however, that people do have different patterns of ability. For example, someone who may possess a great deal of problem-solving ability may find it difficult to relate to patients and colleagues. Therefore, try to evaluate the candidate separately on each factor.

Appendix 17
AMERICAN BOARD OF ORTHOPAEDIC SURGERY

INSTRUCTIONS TO CANDIDATES
Oral Examinations

As indicated on the information sheet sent to you earlier, the oral examination will consist of five one-half hour examinations. Three of these, *Problem-Solving Adult*, *Problem-Solving Children's* and *Problem-Solving Trauma*, will focus on your ability to deal with realistic clinical problems. The fourth, *Simulated Interview*, will focus on your ability to relate effectively to patients and colleagues, and the fifth, *Observation and Interpretation*, will focus on your ability to observe and interpret data. All of the examinations will require you to discuss realistic case material. In the first four examinations the case material will be presented to you in the form of written descriptions; in the last examination you will be required to observe and interpret x-rays and slides. Detailed descriptions of the types of problems to be used and the procedures you will be asked to follow are given below.

Please read these instructions carefully because there will not be time for the examiner to present detailed instructions during the examinations.

THE PROBLEM-SOLVING ADULT, CHILDREN'S AND TRAUMA EXAMINATIONS

General

Four different types of problems will be administered. The *Defense of Therapy Problem* and *Diagnostic Problem* will be administered in the Adult and Children's Examinations. The *Emergency Treatment Problem* and *Complication Problem* will be administered in the Trauma Examination.

Note that while the instructions indicate that you will have a maximum of 13 minutes for each problem, in some cases you may finish a problem earlier.

Defense of Therapy Problem

This type is designed to test your *problem solving ability* in formulating a plan of treatment for a particular patient. You will be given a protocol of a case. This will include a summary of the pertinent historical, physical and laboratory data. Relevant x-rays and information about the diagnosis will be available. If you believe that you need more data before you can formulate a definitive plan of management, you may ask the examiner.

You will be given about 3 minutes to read the protocol. At the end of that time you will be expected to state in your own terms what you feel the problem is, the plan of therapy you would recommend, and your reasons for the treatment you select. You will have about 10 minutes to complete this problem.

The examiner may question you from time to time in order to determine WHY you selected a given approach. He is fully aware that there are many methods to treat a given problem, but he is interested in how you arrived at YOUR approach. He may question you about the entire concept involved in your method or about your ideas behind given parts of your therapy.

Your score on this part of the examination will not depend so much upon the method you select as it will upon your reasoning in deciding upon an approach for this particular patient. (Clinical judgement.)

Complications Problems

This type is designed to test your ability to formulate a plan of management for a patient in which some complication has developed. You will be given a summary of a case which will include relevant historical, physical and laboratory data, including x-rays. The case description will also include the original diagnosis, the steps followed in treatment and the existing complications.

You will be given about 3 minutes to read the protocol. At the end of that time you will be expected to outline a program for the management of the patient. You may ask the examiner to provide information on the results of diagnostic or therapeutic procedures if the plan of management you outline requires such information. You will have about 10 minutes to complete this problem.

The examiner may question you from time to time in order to determine WHY you selected a given procedure. He is fully aware that there are many approaches to such problems, but he is interested in YOUR approach.

Your score on this exercise will depend on the skill you demonstrate in identifying the problem and the reasoning you use in formulating your plan of management.

Emergency Treatment Problem

This problem is designed to evaluate your ability to obtain the critical data necessary to initiate treatment, to diagnose a patient's total problem, ot facilitate management and to institute appropriate initial care for a patient in an emergency situation.

You will be given a brief description of a case as presented in the emergency room. You will have about 3 minutes to study the problem and you will then outline a program for the initial management of the patient - not only with regard to specific actions to be taken, but also to the order and priority of the actions you recommend. You may ask the examiner for specific physical or laboratory findings (including x-rays) in order to make your judgements. The examiner may also indicate the results of each action as you take it. This discussion should take about 10 minutes.

You are reminded that the problem revolves around the initial diagnosis and treatment of the case and has nothing whatever to do with long-term management which will be taken up elsewhere. The cases presented are all taken from the practices of those conducting the examination. You should use the information given as you would in your own practice. Assume that you have available to you the personnel and equipment that you ordinarily would have in you own hospital.

Your score will depend upon your ability to identify the problem effectively and describe an adequate and effective plan of management.

Diagnostic Problem

The Diagnostic Problem is designed to test your ability to gather information concerning a patient and to arrive at reasonable conclusions concerning his illness. You will be given a brief case description including such information as the age, occupation, sex and chief complain of a patient. Your task will be to question the examiner to obtain the necessary information about the history, physical, x-ray and laboratory findings needed to obtain an effective differential diagnosis. You will be given about 10 minutes to gather the information. At the end of this information-gathering session you will be required to present your diagnostic conclusions and your reasons for them. Your explanation of your diagnostic impressions should take about 3 minutes.

Note that during the data-gathering session the examiner will not interpret the data for you, but only give you the same information you might get from a patient. For example, if you ask, "Is there a history of injury?", the examiner will probably say, "Where?", If you then say, "To the arm," the examiner will say, "The patient says that he hurt his arm when he was very small." Therefore, be specific in all your inquiries; the examiner will not volunteer information; you will have to formulate your questions so as to elicit the precise information you want.

If you do not obtain sufficient information from a question, pursue the matter until you are satisfied, but do not waste time exploring what is obviously a "blind alley." Vary your attack should your questioning along one line prove unrevealing. You may allocate your time among various lines of inquiry in any way you prefer, asking additional quesions about any aspect of the investigation at any time during the interview.

Your score on this examination will *not* depend on whether you have arrived at a "correct" diagnosis. It will depend on the skill you demonstrate in gathering information to arrive at a diagnosis, and the logic you employ in explaining your conclusions. It is important, therefore, for you to be sufficiently thorough in exploring reasonable possibliities that ought to be considered in the differential diagnosis. However, as in any real situation, you should avoid wasting your time in asking irrelevant questions or exploring remote and unlikely possibilities. The problems used in this part of the examination are typical of orthopaedic practice in general.

THE SIMULATED INTERVIEW

This examination simulates situations which are quite familiar to you. If you place yourself in the familiar "role" and conduct yourself accordingly you should not experience any difficulty with the format of this examination.

The problems used in these tests are not obscure. They are typical of general orthopaedic practice. There is no attempt to trick you with rare or unusually complex problems. They are designed to give you an opportunity to demonstrate how you would talk with patients and their families about their illnesses and how you would talk with nurses and other physicians concerning patients.

You will be given a description of a clinical situation including diagnosis and the proposed next steps. You will have about 3 minutes to review the case description. You will then have about 6 minutes to interview the examiner who assumes the role of the patient, nurse or another physician. Your task will be to explain to him what problems might have arisen in the management of the patient, what your plans are, and to gain his understanding and cooperation. You should attempt to explain your problem and management in clear, simple terms and gain his confidence. This part of the oral examination will contain three such problems.

Your score will depend upon (1) the effectiveness of the information you provide to meet the requirements of the situation described, and (2) the skill and tact you demonstrate in communicating to the person described.

THE OBSERVATION AND INTERPRETATION EXAMINATION

The ability to observe and interpret accurately is an integral part of the requirements for the successful practice of orthopaedic surgery. The conclusions drawn from direct observation frequently determie the diagnosis, decide the course of treatment or determine the efficacy of treatment. This examination is directed toward the evaluation of your ability to observe and correlate information derived from microscopic slides and x-rays.

After reviewing the material presented, you may be asked to describe what you see, precisely as you might in a written report, indicating any abnormalities that may be present.

After you have finished describing what you see you may be asked to indicate probable causes for any abnormalities you identify.

Your score on this examination will depend mainly upon the accuracy with which you report your observations.

Appendix 18

The American Board of Orthopaedic Surgery
1968 CERTIFICATION EXAMINATION
*Oral Examiner Questionnaire**

On January 18th, following the administration of the oral portion of the 1968 Certification Examination each examiner attending a subject matter debriefing session was asked to complete a questionnaire sampling his reaction to the revised format of the oral examinations. Approximately 220 examiners were employed in administering these examinations, of whom 191 completed the questionnaire. These examiners overwhelmingly endorsed the new oral examinations, although areas for improvement were noted.

The Center for the Study of Medical Education of the University of Illinois tabulated the responses made by the oral examiners to the 21 items in the questionnaire. This tabulation is attached for your reference. It will be my purpose to summarize and interpret this data so that the Board and more specifically the Examinations Committee can make maximum use of the comments and suggestions made and opinions registered.

Interpretive Skills

The examiners in Interpretive Skills found the orientation session least helpful. Had the orientation session dealt more with Interpretive Skills this reaction perhaps would have been different. As it was, two Interpretive Skills examiners attended no orientation session (Item 3).

The assignment given to the Interpretive Skills Examiners was that they should emphasize the question "What do you see?" rather than "What would you do?" How well the examiners followed this like of questioning is not known; however, there is some evidence that they were aware of the restriction this placed on their examination (Item 2).

The expectations of the Interpretive Skills examiners seems to be greater than the performance of the candidates, therefore they tended to be more critical of the candidates, their preparation for the examination and their training programs (Items 4, 10, 19 and 20).

Simulated Interview

There was less agreement that the Simulated Interview provided valuable information about the candidate than about any other section of the examination. This is in part explained by the content of this portion of the examination and in part by the "strangeness" of the examination technique (Items 1 & 2).**

---

* Prepared by L. W. Hattress, Jr., Director, Office of Education and Evaluation, American Board of Orthopaedic Surgery, July 1968

** Also reflected in this conclusion is the feeling that candidates are not being prepared for this portion of the examination (Item 20), to the point where some examiners questioned the fairness of the examination (Item 10).

There is some indication that the addition of x-rays and other illustrative material would improve this portion of the examination (Item 11).

Trauma

The Trauma examiners were most critical of the cases supplied. There were a number of reasons to justify this opinion all of which can be remedied (Item 5). They were also most critical of the x-rays provided which may in part explain why they were critical of the cases (Item 11).

With all, the Trauma examiners felt that the candidates were well prepared to manage trauma problems (Item 20).

Children's Orthopaedics

The Children's Orthopaedic examiners seemed to have some difficulty in applying the 12-point rating scale (Items 7 and 12).

They also seemed to feel more restricted by the standardized format in terms of the questions they could ask than the examiners in other areas (item 8).

Adult Orthopaedics

The Adult Orthopaedic examiners reflected many of the conclusions already reported but in no instance did they feel so strongly. This perhaps is indicative of the more general nature of this examination.

RECOMMENDATIONS

Based on a review of a questionnaire completed by 191 oral examiners following the 1968 Certification Examination it is recommended:

1. That the standardized oral examinations as administered in 1968 be continued as an integral portion of the certification procedure of the Board, but that more attention be paid to the selection of cases and materials.

2. That a person be designated for each subject matter area to serve on the Oral Examination Task Force and that the person be responsible for securing and editing the cases and materials and presenting them orally to the examiners during the orientation session.

3. That orientation sessions be planned in conjunction with meetings of orthopaedic surgeons held in the fall and winter prior to the '69 Examinations and on the day prior to the oral examinations which emphasize;
    - A. The objectives of the oral examination
    - B. The role of the examiner
    - C. The rating of candidates

SUMMARY OF RESULTS OF
EXAMINER QUESTIONNAIRE

American Board of Orthopaedic Surgery
Certification Examination, January 1968

QUESTION		N	Strongly Agree	%	Agree	%	Undecided	%	Disagree	%	Strongly Disagree	%
1. The portion of the examination I administered provided me with valuable information about the candidate's ability in some important area of orthopaedics.	Adult	36	6	17	30	83	0	-	0	-	0	-
	Child	35	4	11	29	83	1	3	1	3	0	-
	Trauma	34	9	26	25	73	0	-	0	-	0	-
	Int. Skills	34	14	41	18	53	2	6	0	-	0	-
	Sim.	50	11	22	28	56	4	8	5	10	2	4
	Total	189	44	23	130	68	7	4	6	3	2	1
2. Unfamiliarity with some of the examining techniques interfered with the candidate's ability to demonstrate his competence.	Adult	36	0	-	7	19	1	3	27	78	1	3
	Child	34	0	-	6	18	3	9	23	68	2	6
	Trauma	34	1	3	2	6	4	12	22	65	5	15
	Int. Skills	34	2	6	5	15	1	3	18	53	8	23
	Sim.	50	1	2	23	46	6	12	19	38	1	2
	Total	188	4	2	43	23	15	8	109	58	17	9
3. The orientation sessions I attended were generally helpful.	Adult	36	11	30	20	55	4	11	1	3	0	-
	Child	36	10	28	25	69	1	3	0	-	0	-
	Trauma	34	12	35	18	53	2	6	2	6	0	-
	Int. Skills	34	5	15	16	47	9	26	3	9	1	3
	Sim.	51	15	29	28	55	2	4	6	12	0	-
	Total	191	53	28	107	56	18	9	12	6	1	.5
4. The candidates' performances were very disappointing to me.	Adult	35	0	-	1	3	4	11	27	77	3	9
	Child	35	1	3	1	3	3	9	26	74	4	11
	Trauma	33	1	3	1	3	1	3	17	52	13	39
	Int. Skills	34	1	3	6	18	2	6	22	65	3	9
	Sim.	51	0	-	1	2	5	10	35	69	10	20
	Total	188	3	2	10	5	15	8	127	67	33	17

QUESTION:		N	SA	%	A	%	U	%	D	%	SD	%
5. I would prefer bringing my own cases because those supplied were poor.	Adult	36	0	-	2	6	7	20	21	58	6	17
	Child	36	1	3	2	6	7	20	21	58	5	14
	Trauma	34	6	18	3	9	3	9	11	32	11	32
	Int. Skills	34	0	-	3	9	2	6	17	50	12	[illegible]
	Sim.	50	1	2	3	6	6	12	26	52	14	28
	Total	190	8	4	13	7	25	13	96	50	48	25
6. The orientation sessions I attended were somewhat confusing.	Adult	34	0	-	2	6	3	9	20	60	9	26
	Child	35	0	-	3	9	0	-	20	57	12	34
	Trauma	34	2	6	3	9	0	-	24	71	5	15
	Int. Skills	32	0	-	2	6	7	22	19	59	4	12
	Sim.	51	0	-	3	6	2	4	33	65	13	25
	Total	186	2	1	13	7	12	6	116	61	43	23
7. The rating system was easy to understand and apply.	Adult	36	5	14	20	55	2	6	9	25	0	-
	Child	36	5	14	14	39	10	28	6	17	1	3
	Trauma	34	7	21	21	62	3	9	3	9	0	-
	Int. Skills	34	5	15	13	38	8	24	7	21	1	3
	Sim.	51	6	12	31	61	7	15	7	14	0	-
	Total	191	28	15	99	51	30	16	32	17	2	1
I would prefer using my own cases because the cases supplied inhibited me from asking important questions.	Adult	36	0	-	4	11	4	11	25	69	3	8
	Child	35	2	6	8	23	1	3	21	60	3	9
	Trauma	34	2	6	2	6	3	9	20	59	7	21
	Int. Skills	34	0	-	2	6	1	3	19	56	12	35
	Sim.	50	0	-	3	6	5	10	31	62	11	22
	Total	189	4	2	19	10	14	7	116	60	36	19
9. I felt ill-at-ease administering this examination.	Adult	36	0	-	1	3	0	-	27	75	8	22
	Child	35	0	-	0	-	1	3	26	74	8	23
	Trauma	34	0	-	2	6	0	-	17	50	15	44
	Int. Skills	34	0	-	0	-	1	3	20	59	13	39
	Sim.	51	0	-	3	6	3	6	37	73	8	16
	Total	190	0	-	6	3	5	3	127	66	52	27
10. My portion of the examination was probably unfair to many candidates	Adult	36	0	-	1	3	1	3	25	70	9	25
	Child	35	0	-	1	3	4	11	23	66	7	20
	Trauma	34	0	-	2	6	1	3	21	62	10	29
	Int. Skills	34	0	-	0	-	3	9	15	44	16	47
	Sim.	50	1	2	1	2	4	8	35	70	9	18
	Total	189	1	.5	5	3	13	7	119	62	51	27

QUESTION:			SA		A		U		D		SD	
		N		%		%		%		%		%
11. The X-rays, slides, and photographs used were clear and easy to read.	Adult	35	4	11	20	57	1	3	10	29	0	-
	Child	36	3	8	19	53	1	3	11	30	2	6
	Trauma	34	2	6	9	26	2	6	14	41	7	21
	Int. Skills	33	10	33	18	60	3	10	2	7	0	-
	Sim.	25	2	8	11	44	7	28	3	12	2	8
	Total	163	21	13	77	47	14	9	40	24	11	7
12. The rating system was generally very difficult to apply.	Adult	35	0	-	6	17	1	3	25	71	3	9
	Child	34	0	-	4	12	7	21	20	59	3	9
	Trauma	32	0	-	0	-	3	9	26	81	3	9
	Int. Skills	34	1	3	6	18	7	21	14	41	6	18
	Sim.	51	0	-	7	14	5	10	34	67	5	10
	Total	186	1	.5	23	12	23	12	119	63	20	11
13. The cases used were generally too easy.	Adult	36	0	-	8	22	1	3	25	69	2	6
	Child	35	0	-	9	26	1	3	23	66	2	6
	Trauma	34	0	-	6	18	2	6	24	71	2	6
	Int. Skills	33	0	-	5	15	4	12	19	58	5	15
	Sim.	49	0	-	8	16	4	8	31	63	6	12
	Total	187	0	-	36	19	12	6	122	65	17	9
14. I am satisfied with my performance as an examiner.	Adult	35	0	-	19	54	13	37	3	9	0	-
	Child	35	1	3	27	77	4	11	3	9	0	-
	Trauma	34	1	3	22	65	11	32	0	-	0	-
	Int. Skills	33	1	3	25	76	5	15	2	6	0	-
	Sim.	49	3	6	29	59	13	26	4	8	0	-
	Total	186	6	3	122	65	46	24	12	6	0	-
15. The cases used were generally too unusual to give me the information I needed about the candidate.	Adult	35	0	-	2	6	2	6	26	74	5	14
	Child	35	0	-	3	9	2	6	27	77	3	9
	Trauma	33	2	6	0	-	0	-	27	82	4	12
	Int. Skills	34	0	-	0	-	0	-	21	62	13	38
	Sim.	49	0	-	5	10	3	6	34	69	9	18
	Total	186	0	-	10	5	7	4	135	72	34	18
16. I liked the idea of standardized case material.	Adult	36	14	39	17	47	3	8	2	6	0	-
	Child	36	15	42	15	42	5	14	0	-	1	3
	Trauma	34	16	47	12	35	3	9	3	9	0	-
	Int. Skills	34	16	47	16	47	2	6	0	-	0	-
	Sim.	51	21	41	24	47	2	4	1	2	3	6
	Total	191	82	43	84	44	15	8	6	3	4	2

QUESTION:		N	SA	%	A	%	U	%	D	%	SD	%
17. I need additional training in the administration of these examinations.	Adult	36	0	-	6	17	2	6	26	72	2	6
	Child	35	0	-	7	20	3	9	24	69	1	3
	Trauma	34	2	6	2	6	7	21	20	59	3	[illegible]
	Int. Skills	34	0	-	6	18	5	15	17	50	6	18
	Sim.	51	1	2	8	16	4	8	30	59	8	16
	Total	190	3	2	29	15	21	11	117	61	20	10
18. The cases used were generally too difficult.	Adult	35	0	-	2	6	0	-	27	77	6	17
	Child	36	0	-	0	-	0	-	31	86	5	14
	Trauma	34	0	-	1	3	1	3	27	79	5	15
	Int. Skills	34	0	-	1	3	0	-	19	56	14	41
	Sim.	51	1	2	1	2	0	-	38	74	11	22
	Total	190	1	.5	5	3	1	.5	142	74	41	21
19. The candidates were, in general, well-educated.	Adult	36	2	6	28	78	3	8	3	8	0	-
	Child	36	3	8	27	75	3	8	3	8	0	-
	Trauma	34	6	18	24	71	3	9	0	-	1	3
	Int. Skills	34	1	3	24	71	5	15	4	12	0	-
	Sim.	50	7	14	36	72	5	10	2	4	0	-
	Total	190	19	10	139	72	19	10	12	6	1	.5
[illegible]0. Most training programs apparently do not adequately train the candidates to take this type of examination.	Adult	34	0	-	8	24	5	15	19	56	2	6
	Child	36	0	-	4	11	10	28	22	61	0	-
	Trauma	34	0	-	5	15	4	12	24	71	1	3
	Int. Skills	34	2	6	7	21	4	12	19	56	2	6
	Sim.	51	1	2	20	39	9	10	19	37	2	4
	Total	189	3	2	44	23	32	17	103	54	7	4
21. I enjoyed administering this examination.	Adult	36	7	20	26	72	3	8	0	-	0	-
	Child	36	9	25	24	66	3	8	0	-	0	-
	Trauma	34	11	32	20	59	3	9	0	-	0	-
	Int. Skills	34	7	21	22	65	3	9	1	3	1	3
	Sim.	51	12	24	30	59	7	14	1	2	1	2
	Total	191	46	24	122	63	19	10	2	1	2	1

Appendix 15

The American Board of Orthopaedic Surgery

1968 CERTIFICATION EXAMINATION

Candidate Questionnaire*

On January 17 and 18, 1968, following each group of oral examinations the candidates attended a debriefing session. At the beginning of this session each candidate was asked to complete a questionnaire giving his opinion about the examinations—written and oral—he had just completed. Processing of these questionnaires was delayed until the results of the examination had been mailed to the candidates so that the opinions expressed would in no way affect the outcome of the examination.

This report is based on the questionnaires completed by 200 candidates selected at random; 100 of whom were successful and 100 unsuccessful in their bid for certification. The data was tabulated by the staff of the Center for the Study of Medical Education (CSME) of the University of Illinois. The results give some indication of the validity of the Board's examinations and are corroborated by the more detailed studies to be reported by CSME.

Of the 200 candidates in the sample, 52 thought they had been successful and were, while 64 had their feelings of failure confirmed, and 36 were not certain about the outcome. Of the remaining candidates, 21 thought they had been successful and were not, while 27 were successful who did not think they would be. Applying these ratios to the actual population taking the examination, it is evident that the majority of the candidates were pleased or surprised by the results while less than 5% thought they had been successful and were not. There will be considerable correspondence with and about this latter group in the months ahead.

On the basis of this questionnaire the unsuccessful candidate may be characterized as being weak in factual knowledge, interpretive skills, and in the application of the knowledge he possesses, but strong in treating trauma and in doctor-doctor and doctor-patient relations. The successful candidate may be characterized as being weak, though not as weak, in factual knowledge, but strong in the treatment of trauma, the diagnosis and treatment of children's disease, in interpretive skills and in doctor-doctor and doctor-patient relations. The most critical difference between success and failure perceived by the candidates appears to be interpretive skills. This has been verified through comparison of the examination results with the ratings submitted on the "Candidate Evaluation Form."

The candidates' opinions were solicited concerning the overall examination process as well as the various portions of the examination. Two thirds of our sample felt that the examinations adequately examined their knowledge and understanding of orthopaedic surgery, while a third thought that the examination contained a great deal of esoteric material. Only 3 candidates felt the examination was too easy and two of them failed, while about 10% felt it was too difficult. Over 80% felt the examination was fair.

---

* Prepared by [illegible], Jr., Director, Office of Education and [illegible]

There was a difference of opinion over the adequacy of the content covered in the examination with the successful candidates feeling that it was not adequate. It should be noted that the successful candidates tended to be critical of the examination while the unsuccessful candidates seemed satisfied with things as they were.

The multiple-choice examination was considered to be the most difficult, irrelevant, ambiguous and unfair portion of the examination. This result was to be expected for the "best answer", forced choice approach is a frustrating experience especially to the highly intelligent.

The patient management problems were considered confusing by a third of the sample but more relevant than the multiple-choice examination and, therefore, fairer and not as difficult.

The trauma oral was considered the least difficult and most relevant portion of the examination. The children's and adult orals followed as a close second and third while Interpretive Skills and the Simulated Interviews were a distant fourth. It is apparent that the candidates would prefer a certification examination made up entirely of oral examinations if given the choice. There is some evidence to indicate that the order of preference for the orals reflects the content of the practice of the orthopaedic surgeon.

The candidates felt that the oral examiners did a good job allowing the candidate to demonstrate his knowledge and ability, to answer questions to the candidate's satisfaction and by consciously working to place the candidate at ease. Only two candidates in our sample felt that examiners were "rude" and both were successful.

Specific questions were asked about the adequacy of the visual aids used in the examination. What criticism there was was directed toward the written examinations, which again was to be expected for, instead of actual X-rays and slides, photographic reproductions were used in these portions of the examination. However, better than 80% of the candidates felt that the X-rays and slides "were clear and easy to read."

On the basis of the response of the candidates immediately following the 1968 Certification Examination we may conclude that the techniques and materials employed to assess competency were relevant and the coverage was adequate. Furthermore the differences between success and failure can in part be explained by different preparation for and perceptions of the examination.

Further work with these questionnaires should be undertaken before broad conclusions are drawn from these findings. Two tasks which should be completed are: (1) the tabulation of the data provided by all the candidates to verify the conclusions drawn from the sample, and (2) an analysis of the data by group and panel to identify any examiner bias* which should be eliminated from future examinations.

SUMMARY OF RESULTS OF
CANDIDATE QUESTIONNAIRE

American Board of Orthopaedic Surgery
Certification Examination, January 1968

QUESTION:	Response	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree	Didn't Answer
1. The examination as a whole had sufficient depth to adequately assess my knowledge and understanding of orthopaedic surgery.	passed	5	56	18	20	0	1
	failed	13	58	14	12	2	1
2. There were too many questions of an esoteric nature.	passed	6	29	21	36	8	0
	failed	5	28	24	33	5	5
3. My training program did a good job of preparing me for this examination.	passed	16	55	20	7	2	0
	failed	19	41	19	17	1	3
4. The examination as a whole seemed very remote from my practice.	passed	3	14	12	62	9	0
	failed	4	15	20	47	10	4
5. The X rays, slides and photographs were clear and easy to read.	passed	38	44	5	12	1	0
	failed	31	50	8	9	1	1
6. Many of the multiple choice questions were very confusing in their wording.	passed	23	36	10	28	3	0
	failed	23	36	8	30	2	1
My unfamiliarity with some of the techniques interfered greatly with my ability to demonstrate	passed	3	12	11	59	15	0
	failed	7	35	14	41	7	3

QUESTION:	Response	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree	Didn't Answer
8. The multiple choice section had too many questions in which two answers could be defended as correct.	passed	45	36	11	7	1	0
	failed	31	42	14	10	1	2
9. The examination was too easy.	passed	0	1	9	49	71	0
	failed	0	2	5	46	43	4
10. The examination as as whole covered all the important topics in orthopaedic surgery.	passed	3	31	17	40	9	0
	failed	7	44	10	31	6	2
11. The technique used in the erasure exercise was confusing to me and I did not get a chance to demonstrate my ability	passed	13	12	6	48	21	0
	failed	16	21	11	34	14	4
12. The techniques used in the oral examination were confusing to me and I did not get a cheange to demonstrate my abilities.	passed	3	7	13	51	26	0
	failed	8	13	13	48	15	3
13. The examination as a whole was fair.	passed	16	67	15	2	0	0
	failed	14	67	10	5	0	4
14. The system of granting my certificate only after I have taken an examination is a good one.	passed	8	45	16	20	9	2
	failed	17	34	21	16	8	4

STATEMENT	Response	Mult. Choice	Erasure	Trauma	Adult	Child	Interpretive Skills	Simulated Interviews
1. Gave me a chance to demonstrate my abilities in some important areas of orthopaedic surgery.	passed	43	61	86	75	82	69	67
	failed	42	52	85	70	75	65	65
2. Most topics covered were important in orthopaedic practice.	passed	52	76	91	85	86	72	78
	failed	52	79	91	84	87	73	77
3. Most topics covered were irrelevant to orthopaedic practice.	passed	29	3	3	4	5	11	9
	failed	20	4	4	7	3	8	3
4. Examination procedures were confusing to me.	passed	5	24	4	5	3	7	12
	failed	11	28	7	12	7	11	10
5. Examination was faire.	passed	52	68	89	83	88	77	76
	failed	64	69	91	88	90	84	85
6. Examination was too easy.	passed	0	1	3	0	0	3	5
	failed	1	5	3	4	3	2	4
7. Examination was too difficult	passed	33	8	3	4	4	5	6
	failed	26	10	3	7	5	17	6
8. The X-rays, slides and photographs were inadequate.	passed	10	15	7	7	8	11	5
	failed	14	13	3	7	4	5	2
9. Examiner was rude to me.	passed	0	0	0	0	0	1	1
	failed	0	0	0	0	0	0	0
10. Examiner was skillful in putting me at ease.	passed	0	0	72	68	74	72	78
	failed	0	0	63	60	66	64	70
11. Examiner did NOT give me a chance to answer questions adequately.	passed	0	0	4	11	3	2	2
	failed	0	0	3	3	1	1	0
12. Examiner gave me ample opportunity to show what I could do.	passed	0	0	77	69	76	73	[illegible]
	failed	0	0	74	72	71	76	74

STATEMENT	Response	Mult. Choice	Erasure	Trauma	Adult	Child	Interpretive Skills	Simulated Interviews
13. Cases used were irrelevant.	passed	0	0	1	3	3	4	..
	failed	0	0	3	5	4	3	1
14. Cases used were central to my practice.	passed	0	0	76	72	69	56	64
	failed	0	0	73	68	63	59	60

		Yes	No	No answer
Do you think you may have failed?	passed	41	32	21
	failed	64	21	15

	Response	Mult. Choice	Erasure	Trauma	Adult	Child	Interpretive Skills	Simulated Interviews
What section?	passed	21	13	4	11	8	8	9
	failed	31	22	8	14	13	24	8

COMMENTS:	Response	Oral	Erasure	General Covera
Positive	passed	38	7	7
	failed	25	5	7

	Response	Oral	Erasure	Mult. Choice	X-rays, sides etc.	Interpretive Skills	Simulated Interviews
Negative	passed	6	11	46	4	7	7
	failed	8	19	30	5	3	7

A few, in each case, commented on largeness of the room anpoor lighting.

Appendix 20

ORAL EXAMINATION

AMERICAN BOARD OF ORTHOPAEDIC SURGERY

Part II: January 1976

Simulated Patient Management Conference

Case Description No. 35

The patient is a 12-year-old Caucasian girl whose family first noted a curve in her back one year ago. There is no family history of scoliosis or other skeletal deformities. The patient herself has no complaints. Her menarche was six months ago.

Four months ago she was seen in the scoliosis clinic of this hospital; physical examination did not suggest any diagnosis other than idiopathic scoliosis (i.e. no cafe-au-lait spots, no positive neurological signs, no leg length discrepancy, and no muscle weakness). There was a right dorsal scoliosis with moderate rotation. The curve improved somewhat with traction and with bending. When she was erect the pelvis was level, and a plumb line dropped from C7 fell in the center of the intergluteal cleft. No other deformities or abnormalities were noted.

X-rays were taken at that time (Series A on the view box). She was scheduled for re-evaluation in three months; at that time (one month ago) she was re-examined (no changes noted in examination) and new X-rays were taken (Series B on the view box).

She was admitted to the hospital for further evaluation; routine laboratory tests are normal. The working diagnosis is idiopathic scoliosis.

Prepared with the assistance of the
Center for the Study of Medical Education
University of Illinois, College of Medicine

Appendix [illegible]

SYSTEM OF CONTENT CLASSIFICATION
OF EXAMINATION MATERIALS

I. PATIENT

- A. Adult
- B. Child
- C. Pertains meaningfully to either

II. DISORDER

- A. Trauma
  - 1. Acute Skeletal
  - 2. Acute Soft Tissue
  - 3. Post-Traumatic
- B. Disease
  - 1. Congenital
    - a. Skeletal
    - b. Neuromuscular (includes all C.P.)
    - c. Other Soft Tissue
  - 2. Inflammatory
  - 3. Neoplastic
  - 4. Metabolic
  - 5. Degenerative
  - 6. Neuromuscular
  - 7. Idiopathic
  - 8. Neurocirculatory
- C. Pertains meaningfully to either
- D. No Disorder

III. PART OF BODY

- A. Upper Extremity
  - 1. Hand
  - 2. Wrist
  - 3. Elbow
  - 4. Forearm
  - 5. Arm
  - 6. Shoulder
- B. Trunk and Head
  - 1. Head
  - 2. Cervical Spine
  - 3. Thoracic Spine

III. PART OF BODY (continued)

  - 4. Chest
  - 5. Lumbar Spine
  - 6. Pelvis
- C. Lower Extremity
  - 1. Hip
  - 2. Thigh
  - 3. Knee
  - 4. Leg
  - 5. Ankle
  - 6. Foot
- D. Body as a Whole
- E. Non-Applicable

IV. BASIC SCIENCE

- A. Anatomy
  - 1. Functional
  - 2. Gross
- B. Physiology and Biochemistry
- C. Pathology
- D. Biomechanics
- E. Other
- F. Non-Applicable

V. CLINICAL

- A. Diagnostic
- B. Treatment
  - 1. Non-Operative (includes manipulation)
  - 2. Operative (must draw blood)
- C. Post-Treatment
  - 1. Complications
  - 2. Rehabilitation
  - 3. Orthotics
  - 4. Prosthetics
- D. Non-Applicable

Appendix 22

American Board of Orthopaedic Surgery
Examination Blueprint

COGNITIVE DOMAIN

PROCESS

- I. Knowledge (20%)
  - A. Recall (10%)
  - B. Recognition (10%)
- II. Application (80%)
  - A. Interpretation & Analysis (20%)
  - B. Problem solving(30%) (diagn. & manag.)
  - C. Judgment (30%)*

CONTENT

Dimension		
I Type of Patient	A. Adult (75%)	B. Child (25%)
II Type of Disorder	A. Trauma (40%) 1. Acute skeletal (20%) 2. Acute soft tissue (10%) 3. Post-traumatic (10%) C. No Disorder	B. Disease (60%) 1. Congenital (10%) a. Skeletal (4%) b. Neuromuscular (4%) c. Other soft tissue(2%) 2. Inflammatory (10%) 3. Neoplastic (10%) 4. Metabolic (10%) 5. Degenerative (10%) 6. Miscellaneous (10%)
III Parts of the Body	A. Upper extremity (30%) 1. Hand (5%) 2. Wrist (5%) 3. Forearm (5%) 4. Elbow (5%) 5. Arm (5%) 6. Shoulder (5%) B. Trunk and Head (30%) 1. Head (5%) 2. Cervical spine (5%) 3. Thoracic spine (5%) 4. Lumbar spine (5%) 5. Chest (5%) 6. Pelvis (5%) 7. Other (2%)	C. Lower extremity (30%) 1. Hip (5%) 2. Thigh (5%) 3. Knee (5%) 4. Leg (5%) 5. Ankle (5%) 5. Foot (5%) D. Body as a whole (10%)
IV Type of Information	A. Basic Science (35%) 1. Anatomy (10%) 2. Physiology (5%) 3. Biochemistry (5%) 4. Pathology (5%) 5. Biomechanics (5%) 6. Microbiology (5%) 7. Other	B. Clinical (65%) 1. History (15%) 2. Physical findings (15%) 3. Diagnosis (20%) a. Radiograph (8%) b. Laboratory (4%) c. Special studies (4%) d. Evaluation of impairment (4%) 4. Prognosis (15%)
V Type of Management	A. Nonoperative (35%) B. Pre-operative (15%) C. Operative (30%)	D. Postoperative (20%) E. Rehabilitation (5%) F. Does not relate to management

* effective and [illegible] use of available information and weighing of the various factors involved

AFFECTIVE DOMAIN

		CONTENT									
PROCESS		I. Relationships	A. Patients	B. Colleagues	C. Community	D. Profession	II. Settings	A. Clinical	B. Educational	C. Research	D. Other
Ability to											
I. Relate effectively to patients	(3)*										
A. Show insight into personal relationships (motivation, etc)	(3)										
B. Empathize	(1)										
C. Direct and control interviews	(3)										
D. Respect the privacy and dignity of patients	(1)										
E. Show concern and consideration	(3)										
F. Inspire confidence	(2)										
G. Demonstrate awareness and understanding of socio-economic background	(3)										
II. Relate effectively to colleagues (including willingness to seek consultation)	(3)										
III. Express and accept disagreement in a constructive fashion	(3)										
IV. Think and act in an unselfish fashion	(1)										
A. Avoid prejudice concerning colleagues and medical problems	(2)										
B. Put patient's needs above personal gratifications	(1)										
V. Desist from treatment when appropriate	(2)										
VI. Improvise and demonstrate receptiveness to new ideas	(2)										
VII. Recognize and accept the responsibility for own errors and limitations	(1)										
VIII. Be thorough and persistent in appropriate situations	(1)										
IX. React effectively to emergency situations	(1)										
X. Organize one's work effectively including coordinating the work of others	(3)										
Willingness to											
XI. Recognize and accept the responsibilities of physicians to his patient's colleagues and community	(2)										
XII. Accept the responsibility to develop his own medical knowledge	(1)										
XIII. Display integrity	(1)										

*The numbers in the parentheses indicate the priorities established by the Sub-committee on the Affective Domain
1 = Highest, 2 = Moderate, 3 = Lowest

PSYCHOMOTOR DOMAIN

PROCESS		CONTENT I. Physical Examination	II. Surgery	III. Rehabilitative Procedures	IV. Medical Emergency	V. Confrontations with patients, colleagues, students
I. Manipulative skills						
A. Manual and finger dexterity (Skillful, controlled purposeful and efficient use of the fingers and hand)	(2)*					
B. Fine psychomotor coordination (Ability to do highly controlled adjustments involving the entire arm)	(2)					
C. Multiple limb coordination (Ability to coordinate all extremities in performing movements)	(3)					
II. Spacial relationships (Ability to comprehend and perform effectively in three dimensions)	(2)					
III. Ability to apply mechanical principles to surgical procedures	(1)					
IV. Ability to apply cognitive knowledge of living anatomy in the performance of surgical and physical examination procedures	(1)					
V. Physical control under emotional distress	(1)					
VI. Communication skills (verbal and nonverbal)	(1)					

* The numbers in parentheses indicate the priorities established by the Subcommittee on the Psychomotor Domain.
1 = Highest, 2 = Moderate, 3 = Least

AMERICAN ACADEMY OF ORTHOPAEDIC SURGEONS

[illegible]

February 3, 4, 5, 196[illegible] Chicago, Illinois

Participants:

Examination Committee, American Board of Orthopaedic Surgery

Dr. Charles E. Gregory, Chairman 1]
Dr. [illegible] 3]
Dr. [illegible] 2]
Dr. [illegible] 2]
Dr. [illegible] 2]
Dr. [illegible] 1]
Dr. William A. [illegible] 3]

Consultant, American Board of Orthopaedic Surgery

Dr. Lee [illegible] 3]

[illegible]

Dr. George E. Miller 3]
Miss [illegible] 1]
Mr. [illegible] 2]
Dr. [illegible] 3]

Graduate Education Committee, Subcommittee on [illegible], American Academy of Orthopaedic Surgeons

Dr. William E. [illegible]

1] Subcommittee on the present examination of the Orthopaedic Boards
2] Subcommittee on the revised examination of the Orthopaedic Boards
3] Subcommittee on the Psychomotor and Affective Domains

Appendix 23

University of Illinois

College of Medicine

Center for the Study of Medical Education

ORTHOPAEDIC TRAINING STUDY

Working Papers for: Special Meeting of Examination Committee

May 6-8, 1966

INTRODUCTION

The major purpose of this special meeting is to prepare a blueprint or set of specifications for the certification process as conducted by the Board and to outline procedures to be followed in determining the rate and methods of implementing the blueprint. In this connection it should be noted that in developing the blueprint we should try to design an "ideal" one that will serve as a long-term guide to policy. Its practical implications in regard to present practice and the feasibility of change should not enter into consideration at this stage; only after the guide has been completed and approved should its implications for format and its application to a specific forthcoming examination be considered.

These working papers have been prepared to familiarize members of the Examination Committee with principles that are commonly followed in the development of test specifications. The illustrations provided in the text are merely suggestive; clearly the nature of the categories and the extent of detail in any blueprint must be determined by those familiar with professional requirements in the field under consideration.

5

THE PURPOSES OF AN EXAMINATION BLUEPRINT

The primary function of a blueprint is to enable a policy-making group to maintain effective, systematic control of the nature and content of an examination, while delegating the construction and assembly of examination materials to others. The time gained through such delegation and decentralization enables the parent group to concentrate on review and revision of the examination as a whole to ensure that it meets specified purposes. Without this delegation such a group generally finds itself so caught up in operational details that there is little time to establish and review overall policy and procedure.

In short the blueprint or specifications for an examination are intended to serve the same functions in selecting and assembling examination formats and materials as does the architect's blueprint in selecting and assembling building materials. Thus, an effective blueprint will stipulate in detail the categories to be sampled and the weight of each category in the total examination. As will be seen later the number of categories that should be stipulated is far greater than would result from a classification of broad subject areas. On the basis of the decisions concerning classifications responsible sub-groups can then prepare cases, questions, etc., of the type and in the quantities required to meet the specifications with respect to both content and process. These materials can then be stored and retrieved in a logical fashion according to the specifications. From such a library one individual can then assemble an examination for review and revision by the full examination committee.

After the examination is administered, the results can be reported in terms of the categories of the blueprint; profiles of individuals and types of training can be prepared; and progress can be charted from year to year. In addition, new examination committees will have the benefit of established guidelines as well as a point of reference for the review and updating of the examination in the light of progress in orthopaedics and in the science of examining.

STEPS IN PREPARING A BLUEPRINT

Tests are thought of by evaluation specialists as work samples from which decisions are made concerning the capabilities of the persons being tested. For example, physicians do a number of things: they take historys, they set broken bones, they teach residents, they make diagnoses, they decide on courses of treatment, they reassure anxious parents, they order X-rays. Some tests are designed to be *direct* samples of the performance of these tasks; most are *indirect*. For example, a physician presumably needs to know the anatomy of the back in order to operate on a slipped disk. Thus a question requiring the candidate to recall anatomical information about the back is a *direct* sample of his ability to recall information but at best it is only an *indirect* measure of his operating skill. Observation of the candidate actually performing the operation would be a *direct* measure of this skill.

Whether the tasks finally included in an examination are direct or indirect measures of performance, it is useful in developing a blueprint to focus on the idea that, essentially, they are samples of behavior or performance.

The Concept of Domain

Almost everything people do can be divided into three broad performance categories or domains of behavior: the cognitive, the affective and the psychomotor. This concept of domain provides a useful starting point in developing a blueprint.

Those tasks which require predominantly intellectual skills are placed in the *cognitive* domain: e.g., recalling information, using information to solve problems, weighing several factors to arrive at a decision, predicting an occurrence on the basis of specific information. Most written tests or certifying examinations sample this domain almost exclusively.

The *affective* domain includes those tasks in which feelings and attitudes predominate: e.g., relating to patients, demonstrating integrity, relating to colleagues, manifesting self-control in emergencies. These are the things most often included in the term professional behavior. Some types of oral tests and interviews are designed to probe these qualities; at present, in the certification procedures of the Board, evidence on this domai is collected almost exclusively by the Committee on Eligibility.

The psychomotor domain refers to those activities in which manipulative skill is of primary importance: e.g., performing surgery, setting broken bones, applying splints. Rating forms for use in observation are often employed to measure performance in this domain.

It is recognized that many tasks overlap all three domains. For example, in performing a physical examination of a severely injured but conscious patient the orthopaedist is operating at a high level in all three domains. For purposes of designing a blueprint, however, it has been found useful to analyze and specify the domains separately.

The Two-Way Grid

The classification of tasks into domains is only the beginning. A task within any domain must be further specified with respect to both the nature of the process and of the content which it samples. The critical incident study outlines 92 different areas (e.g., skillful gathering of information, developing diagnoses, exercising judgment in deciding on care, exercising skill in operative procedures, etc.). These tasks are all processes, but they do not exist in a vacuum. A man must be skillful in gathering information about something and must decide on the care of something. A physician can go through a process in treating some patients which an objective observer would interpret as showing "good judgment" but the same physician could go through a similar process in treating other patients that the same observer would label as showing "poor judgment." Before decisions can be made about whether a physician usually demonstrates good judgment it is necessary to obtain samples of his work which provide evidence about his judgment in treating specific conditions in various types of patients. These specifics represent the content elements of the tasks.

In a blueprint these two elements of work samples or test questions (content and process) are usually shown in the form of a two-dimensional chart or grid. Content factors are conventionally placed on the vertical axis of the grid and process factors on the horizontal axis. Any specific test question is located on this chart in terms of the two axes. Figure I is an example of such a grid in the cognitive domain.

Figure I

SAMPLE TWO-WAY GRID

Cognitive Domain

PROCESS CONTENT	Judgment in Deciding on Care	Skill in Gathering Information	Ability to Develop a Diagnosis	ETC.
Children's Orthopaedics				
Adult Orthopaedics				
ETC.				

The Concept of Dimension

To illustrate the further subdivision of each axis consider the following question from the May 1964 Part I Orthopaedic Examination.

> HISTORY
>
> A 4-year-old boy had the onset of pain in his right knee three days after a fall in which he suffered abrasions of his face. The patient developed fever and swelling of his right knee. The patient was treated with antibiotics for three weeks. Although the patient's fever was controlled, the patient continued to have swelling of his knee and restriction of motion. Six weeks after the onset the patient had a 45° flexion contracture of his knee, tenderness over the medial femoral condyle and parapatellar swelling. See the X-ray of the knee and a histologic section of tissue from the lesion.
>
> The lesion is:
>
> a) traumatic.
> b) benign neoplasm.
> c) inflammatory.
> d) malignant neoplasm.

With respect to process the sample question requires the candidate (1) to read and interpret X-rays and histologic data, (2) to interpret clinical information and (3) to use the information (both verbal and visual) in arriving at a diagnosis. Such activities should therefore be specified along the process axis.

With respect to content this question appears to deal with (1) a child, (2) some sort of inflammatory disease, and (3) pathology of the knee. However, if the content axis contained such topics as (1) children, (2) disease, (3) knee, etc., it would be difficult to know where to classify the question since clearly it involves knowledge about all three. Consequently, on the content axis, at least, it is necessary to utilize several dimensions, each of which contains mutually exclusive subcategories. For example, some of these dimensions and subcategories might be organized as follows:

Dimension	Sub-Category
Type of patient	Child Adult
Nature of disease process	Trauma Inflammatory disease Neoplastic disease Etc.
Part of Body	Hand Back Knee Etc.

The sample blueprint in Figure II suggests how the resulting blueprint might look.

Once the dimension and sub-categories have been fully specified in each domain it is necessary to assign weights to each in order to assure that any given examination is an appropriately balanced sample of the content and processes which determine competence. As Figure II indicates the weights for the sub-categories in each dimension should add up to 100%. If the arbitrary weight shown on the sample were followed, 60% of the questions would deal with clinical problems; 70% of these clinical problems would deal with trauma. It would NOT be necessary, however, for the process categories to be equally represented in each sub-category of content; thus in the example there is no implication that 20% of the trauma questions should sample recall; in this illustration it would be necessary merely to assure that overall about 20% of the questions in the total examination be recall. If the trauma questions were mostly recall then a higher proportion of questions in other categories would need to entail problem-solving.

THE CHARACTERISTICS OF AN IDEAL BLUEPRINT

Figure II may look rather complex and unwieldy but the process of creating and using such a grid is usually not a difficult one. Most committees have found that the task proceeds rapidly, once the ground rules are agreed upon. In setting these ground rules the essential characteristics of a useful blueprint must be considered:

The first is completeness. This means that as far as possible the blueprint should cover all the important areas of concern.

The second important characteristic of a blueprint is that the topics included in each dimension and sub-category be mutually exclusive. To illustrate the problem, the categories infant, child, adult, geriatric patient do not meet this criterion and because they do not they would create difficulties in classification.

The third important characteristic of a blueprint is that the categories be meaningful and easy to apply. They should be categories that are clear to authors of questions, examinees, the Board and Training directors, and they should be ones that all would recognize as relevant.

THE RELATION BETWEEN THE BLUEPRINT AND PRESENT PRACTICE

It should be emphasized that a chart of the type shown in Figure II is more of a master plan than a set of operational specifications for a specific examination. When a blueprint is first established it is often impossible to put the entire plan in operation immediately because of expense or the imperfect state of the science of examining, but it is extremely important if the blueprint is to be a guideline both now and in the future to design it as though the ideal could be attained. Skills in constructing and interpreting tests are improving at an impressive rate, and a blueprint of the ideal will serve to indicate the direction that experimentation must take to improve the certification process.

It is to be expected, however, that much of the blueprint will have immediate applicability and will facilitate the delegation of specific responsibilities for the construction and assembly of examination materials, and for their storage and retrieval.

PRE-MEETING TASKS

In order to expedite the business of the meeting it is requested that two tasks be completed by each participant before the meeting.

1. Review the enclosed samples of evaluation techniques now used by the Board to determine how each might be categorized in terms of content and process, and therefore what types of dimensions and sub-categories will probably be needed on each axis of the two-way grid.

2. Assign priorities to the process categories included in the critical incident study (see procedure below) and return the information requested to Mr. Levine by May 3rd.

Procedure for Assigning Priorities

A strong case could be made that all the categories listed in the critical incident study are highly important, but it is understood that individuals are stronger in some areas than others. Assume that you had to recommend ONE orthopaedist to serve as a model of the competent orthopaedist in private practice. In weighing the strengths and weaknesses of the candidates for this award, some categories would be more important than others. Use this criterion in assigning priorities.

FIRST: Each of the behaviors in the critical incident study has been typed on a separate card. Look at each card and place it in one of three piles. Three header cards have been provided for labeling these three piles. Pile A should contain those categories you consider MOST IMPORTANT. Pile B should contain those categories you consider of MODERATE IMPORTANCE and Pile C should contain those categories you consider as LEAST IMPORTANT. The three piles should contain about equal numbers of cards. If there is an imbalance Pile B should contain more cards than A or C.

SECOND: Take Pile A and divide into two piles. A1 should contain the MORE IMPORTANT behaviors and A2 those of LESSER IMPORTANCE. The two piles should be of about equal size but in case of an imbalance A2 should have more cards than A1.

THIRD: Take Pile C and divide it into two piles. C1 should contain the MORE IMPORTANT behaviors and C2 should contain those which are of LESSER IMPORTANCE. The piles should be of about equal size but in case of an imbalance C1 should have more cards than C2.

FOURTH: Place the header cards A1, A2, B, C1, C2 on TOP of the appropriate piles; put a rubber band around each pile.

FIFTH: Mail the cards in the enclosed envelope before May 2nd.

Figure II

SAMPLE BLUEPRINT

				PROCESS								
				Cognitive Domain			Affective Domain			Psychomotor Domain		
	Dimensions	Sub-Categories	Weights	Recall	Problem Solving	Etc.	Ability to Relate	Integrity	Etc.	Operative Skill	Skill in Applying Prosthetic Devices	Etc.
I	Parts of body	Hips	10%	20%	40%	40%	20%	60%	20%	70%	25%	5%
		Back	25%				X	X	X			
		Etc.	65%				X	X	X			
II	Type of patient	Adult	50%				X	X	X			
		Children	50%				X	X	X			
III	Type of complaint	Trauma	70%									
		Inflammatory disease	20%									
		Etc.	10%									
IV	Subject	Clinical	60%									
		Basic Science	40%									
V	Types of individuals	Patients	30%	X	X	X				X	X	X
		Colleagues	40%	X	X	X				X	X	X
		Students	30%	X	X	X				X	X	X

Appendix 24

Working Papers--TASK FORCE ON ORAL EXAMINATIONS

I. INTRODUCTION

The Orthopaedic Training Study was undertaken 2-1/2 years ago as a joint project of the Center for the Study of Medical Education and the American Board of Orthopaedic Surgery. The purpose of the study was to analyze the development of competence in orthopaedic surgery in order to see if certification could be based on performance rather than on the passage of some arbitrary length of time. The first stage of the study required the development of a behaviorial description of the components of competence that characterize the effective orthopaedist. The second stage required the analysis of the techniques of evaluation used by the Board when the study was started. The third stage required that orthopaedic surgeons and evaluation specialists work in concert to devise new procedures and improve the old. The study is now in the middle of this third stage.

During the past eight months a number of task forces have been assembled to assist in this process of developing new evaluation procedures and reviewing and revising the older ones. The oral examinations of the Board represent one of the most important of its evaluation procedures. They require a very heavy investment of time and energy on the part of the orthopaedic profession. In January 1966, for example, over 150 examiners spent two days examining approximately 400 candidates--a commitment of over 4,000 man hours. It is important that this time be of maximum benefit to the training and certification of orthopaedic surgeons.

II. THE PURPOSE OF THE TASK FORCE

Over the past 2-1/2 years, there has been a considerable amount of research on the effectiveness of the present oral examination procedures of the Board, and the beginnings of innovation in these procedures. The mission of this Task Force is to review the research, the innovations already made, and the innovations that have been suggested, and make recommendations to the research staff for further research and to the Board for implementation of the findings produced so far by the study.In addition it is hoped that the Task Force will assist in carrying out the research.

to Chicago and given a two-day training session at which they discussed and practiced the administration and rating of the new examinations.

The new examinations were tried out and analyzed during the January 1966 Part II examinations. The results of this administration were as follows: for the most part the Patient Interviews met acceptance by both the examiners and the candidates. Both were somewhat puzzled about the rating of the conference examination. (2). The Patient Interviews met reasonable standards of rater reliability for tests of their type; the conference did not; (3) The new orals did not correlate very highly with other evaluative techniques; (4) Different teams of examiners used different standards (some were easier than others) on the Patient Interviews. This was also true of the traditional oral examinations.

This study raised some questions about the validity and reliability of oral techniques which could not be answered from the data available. A fourth study was therefore inaugurated in the fall of 1965 to answer these questions. In this study the Patient Interviews and the traditional Adult Orthopaedic examination were administered to 236 residents at all levels of training in five different areas in the United States. Provision was made to obtain reliability data not only on two raters reviewing the same examination but also on two examiners asking different sets of questions using the same type of examination. These data are still being processed. Preliminary review of some data indicates that both the new and the traditional oral examinations are much less reliable than written examinations.

IV. EVALUATING BOTH THE NEW AND TRADITIONAL ORAL EXAMINATIONS

Evaluation specialists have developed certain criteria for evaluating testing techniques. The results of these studies can best be interpreted and applied if they are reviewed in terms of the five criteria: (1) relevance, (2) validity, (3) reliability, (4) efficiency, and (5) effect on the educational program. Each of these is discussed below:

A. Relevance

A good examination is one which samples areas of competence directly related to the purposes of the examination. Since the purpose of the Orthopaedic Board examination is to certify competency in the practice of orthopaedics, the oral examinations should sample these areas of competence. The new oral examinations were especially designed to sample

III. DESCRIPTION OF RESEARCH

There have been four studies so far which have yielded important information which bears upon the oral examinations conducted by the American Board of Orthopaedic Surgery.

The first of these was the aforementioned critical incident study. This study yielded a behavioral description of competence in orthopaedic surgery which covered 94 categories. (See DocumentI) The examination committee of the Board then used these descriptions as a guide in developing a set of tentative examination specifications or blue print which would serve as guide to the development of future examinations. (See Document II.) The importance of this study to the oral examinations is that certain behaviors such as "ability to relate to patients" and "ability to relate to colleagues" which are identified as components of effective performance by the critical incident study can most readily be assessed by some type of oral examination.

The second research study was an analysis of the traditional oral examinations used by the Board. In January 1965 a research team consisting of five physicians and three educators observed 144 of the 2,000 individual examinations conducted by the Board. The team recorded 6,868 questions during their observations and categorized the type of competence being assessed by each. The data indicated that these questions for the most part measured the candidates ability to recall (rapidly and under stress) isolated fragments of information.

As a result of these studies, the project staff developed three new types of orals which attempt to measure some areas of competence listed in the critical incident study which were not being measured by other techniques. These examinations were: (1) The Simulated Patient Diagnostic Interview which requires a candidate to play the role of physician and elicit information from an examiner who plays the role of a patient during the history taking session and who also provides information on the physical examination and laboratory studies. At the end of this information session the examinee must explain his diagnostic impressions to the examiner or examiners. (2). The Simulated Proposed Treatment Interview which requires a candidate as a simulated physician to explain a proposed treatment to the examiner as a simulated patient (3). The Simulated Patient Management Conference in which a group of five candidates simulate a conference at which they discuss the treatment of several cases. In order to administer these examinations successfully, approximately 35 examiners were brought such areas so it would be hard to question their relevance, although some might claim that the areas of competence sampled by the new orals were less important than some other areas.

If the content analysis of the traditional orals is accurate, then much of the material covered by these examinations is open to the charge of lack of relevance since it has often been found that recall of isolated facts may have little relationship with the ability to use information to solve problems. It is probable that some oral examiners use questions which are much more relevant than others.

B. Validity

It is not sufficient to state that a test measures a particular type of ability. There must be evidence that it does so. It is not always easy to obtain such evidence.

One way to approach the problem of validity is to see if the content of the test coincides with the test's stated purpose. The new oral tests simulate typical situations which require the abilities evaluated by the test. It may be, however, that the simulation is not close enough to the situation being simulated to provide adequate information on the candidate's competence. The traditional orals were designed to gain some information on the candidate's understanding of various important concepts in orthopaedics. And most of the questions are selected with that purpose in mind. The possibility exists, however, that the evaluate recall rather than understanding.

Another way to test the validity of a test is to see if candidates perform on the test the way they perform in actual situations when observed over a long period of time by competent observers. Such information is now being gathered. It must be emphasized that there are some areas of competence evaluated on tests that qualified observers seldom have the opportunity to evaluate.

A third way to measure test validity is to see if the test results agree with some hypothesis concerning the areas of competence being measured. For example, residents beginning their training should do less well than residents ending their training. Such data is now being gathered, and it is hoped that a preliminary report can be presented to the Task Force.

C. Reliability

Reliability has to do with the consistency which a particular test measures a particular factor. The reliabilit of a test is greatly affected by errors caused by disagreements among raters and by lack of adequate sampling of the behavior being evaluated. Oral examinations are inherently less reliable than objective written tests, and there is evidence that both the new and traditional orals are relatively unreliable. While this unreliability does not obviate the use of such tests, it does indicate that they are best used in combination with other measures of competence as part of test batteries. It also indicates the importance of training examiners and standardizing case materials and procedures to minimize this unreliability.

D. Efficiency

It is obviously impossible to test all areas of competence with complete reliability. It is therefore important to use resources wisely and to avoid concentrating all the testing in one area while leaving other areas unsampled.

The content analysis seems to indicate that this is the most serious failing of the traditional orals. They seem to duplicate the same factors as the multiple choice examinations. The new orals are an attempt to measure some additional areas of competence not evaluated by other examinations.

E. Effect on the Educational Program

Examinations have an obvious effect on the attitudes of students and on the things they study. If an examination covers most of the important areas of competence, the student's attention will be directed toward these areas. While it is desirable to focus the attention of residents on content to facilitate their review of the important concepts in a field, too much emphasis on content without corresponding attention to process will encourage rote memorization. Little attention will be paid to understanding or application and less to the non-cognitive aspects of effective performance.

V. SUGGESTIONS BY THE STAFF OF THE ORTHOPAEDIC TRAINING STUDY

As a result of the studies and analysis above, the following suggestions were made by the staff of the Orthopaedic Training Study.

A. Standard procedures for all oral examinations should be adopted. This implies the development of detailed instructions for conducting the examination, and standardized rating forms and case materials for the conventional examinations analogous to those used for the new orals.

B. Systematic training programs for oral examiners and question authors should be instituted to insure continual updating of the question pool and renewal of the pool of experienced examiners for both oral and written tests.

C. Consideration should be given to alternative methods for establishing passing grades at the requisite level of competence rather than by a "curve" or some other purely statistical criterion. This may very well require reporting and judging in terms of profiles of performance rather than simple summations or averages. The final judgment on this point will be made by the Task Force on Scoring but this Task Force should be aware of the possibility.

In addition to the above suggestions the Center has continued its research into oral examining techniques. The following new techniques have been explored for consideration for introduction into the oral examining process.

A. An oral examination of the cognitive aspects of surgical skill. A copy of a tentative rating form and description of the examination is enclosed. (See Document III.)

B. An oral examination focusing on the ability of an examinee to select and defend therapy. A copy of the rating form and description of the examination is enclosed. (See Document IV.)

C. An oral examination used to analyze the ability of the examinees to observe visual materials.

VI. LOOKING TO THE FUTURE

The suggestions made earlier are relatively easy to implement within the present structure of the oral examinations. It may be, however, that the whole present organization of oral examinations might need to be revised. The multiple choice examination is well designed to measure a wide sample of knowledge.

To a more limited extent it can be used to measure problem solving and observational skills. The erasure technique is designed to measure problem solving and some aspects of clinical judgment. In the interests of efficiency the oral examination should concentrate on non-cognitive skills such as ability to relate to patients and higher level problem solving.

Instead of being tested in five subject areas the examinee could be required to demonstrate various abilities. Some of these abilities are listed in the blue print:

1. ability to relate effectively to patients
2. ability to relate effectively to colleagues
3. ability to accept and express disagreement when appropriate
4. ability to think and act in an unselfish fashion
5. ability to desist from treatment when appropriate
6. ability to be thorough and persistent in appropriate situations
7. ability to react effectively in emergency situations
8. ability to organize one's work effectively including coordinating the work of others.

In addition the tests could focus on some of the higher level problem solving abilities such as

1. ability to apply basic science information to the solution of clinical problems.
2. ability to select and use the appropriate information in solving clinical problems
3. ability to indicate the appropriate criteria for making clinical judgment

These abilities could be tested by having the candidates react to various clinical problems which would be central to the practice of orthopaedics. The ratings would therefore be based primarily on the candidate's approach to the problem and not on his supply of information. Of course if he lacked information on these central problems he would probably fail.

Some typical problems might be:

1. to explain a treatment to a patient
2. to indicate rationale for treatment
3. to analyze a problem using information on pathology
4. to consult with colleagues on treatment
5. to react to problems in medical ethics

6. to obtain history from a patient
7. to indicate the important parameters in solving a clinical problem
8. to indicate the anatomical problems in adopting a particular surgical approach
9. to indicate how one might react to solving an impossible treatment problem

These suggestions present many problems in implementation. However, they seem to point the way to the most efficient use of the oral examination technique.

Appendix 25

Working Papers: TASK FORCE ON WEIGHTING AND SCORING

I. Responsibility of the Task Force on Weighting and Scoring

Each of the other Task Forces appointed by the Board has necessarily dealt with only a single aspect of the certifying process but their combined recommendations have important implications for the weighting and scoring of the various types of data obtained for purposes of certification. It is the responsibility of this Task Force to review the present procedures of the Board, the research results obtained to date from the Orthopaedic Training Study and the activities of the other Task Forces with a view to making recommendations for needed revision in the scoring and weighting of the various components of the certification process.

These working papers are designed to assist the Task Force in analyzing the issues in weighting and scoring and in developing sound procedures for dealing with this perennial problem.

II. Analysis of the Problem

In arriving at a set of recommendations regarding weighting and scoring three issues must be resolved:

1. What scores should be obtained?
2. How should these separate scores be weighted and combined?
3. How should the standard of satisfactory performance be established?

In considering what scores should be obtained it is important to note that at present independent scores are obtained on each type of examination (i.e., the written and each of five orals) and decisions are made on the basis of these separate scores. This procedure has little except convenience to recommend it. Logically, scores should be obtained on each different aspect of competence rather than each test technique. The work of the other Task Forces suggests the competence can best be described in terms of such processes as the following:

1. Ability to recall factual information relevant to medical (especially orthopaedic) practice.

2. Ability to interpret information.

3. Ability to solve problems in diagnosis and treatment.

4. Ability to relate to patients.

In considering how to weight and combine scores two aspects of the problem should be noted: First, it is necessary to decide whether to treat each score independently or to combine certain scores in order to obtain the most reliable data on which to make a decision about a candidate; secondly, if it is decided to combine certain scores, it is then necessary to decide which scores should be grouped and how much weight should be given to each. At present only the scores on the various parts of the written examination are combined and the weight assigned to each does not necessarily reflect the emphases recommended in the recently developed blueprints.

In considering the problem of setting the passing level on the scores that are obtained it should be noted that traditionally two quite different approaches have been taken: (a) that of setting a conventional percent of the examination as passing (e.g. "75" is passing in the oral) and that of failing a certain percent of the candidates (e.g. the lowest one-sixth of the candidates fail the written). It should be clear that both of these approaches are arbitrary and each has certain inherent difficulties: in the first, "75" may or may not represent "competence" and variations in the difficulty of the examination will produce unintended variations in the standard; in the second, there is no reason to assume in advance that precisely one-sixth and only one-sixth of the candidates are incompetent, and variation in the candidate pool from year to year will produce unintended variation in the standards. While it is not difficult to obtain reliable data to rank candidates on some aspects of competence, it is exceedingly difficult to decide how and where to draw the line between pass and fail.

III. Analysis of the Present Board Procedures

A. Scores Obtained

In the 1966 Part II Board examination eight sets of scores were obtained; score on the Multiple Choice test, the score on the

Patient Management Problems, score on short answer questions, and the scores on the five oral examination. The PMP, Short Answer and Multiple Choice scores were combined into a single score on the written test, but the score on each of the oral tests was considered independently. This structuring of the examination seems to have resulted from the feeling that the written and oral examinations somehow measure different processes; however, serious doubt was cast on this assumption by the results of the content and process analysis of the 1965 examinations. Under these circumstances only considerations of convenience would justify perpetuating the present practice.

B. Weighting

In the present examinations the overall score on the written and five separate scores on the orals are used independently to identify failing candidates and the results on each of these six components are reported separately to candidates. From the point of view of its reliability the overall score on the written has much to recommend it since the variance due to error is minimized and it is therefore farily certain that most of the low scores actually performed less well than most of the passing candidates. However, the decision regarding the weights to be assigned each part of the written examination is fairly arbitrary in the absence of clear guide lines defining the relative importance of the various aspects of competence sampled in each part of the written.

The same principle which indicates the advantages to be gained from combining scores on the various parts of the written examination casts doubt on the practice of independent consideration of each score on the separate parts of the oral. First, it is well known that different teams of examiners may use different standards (see Document I) and thus each score by itself is quite unreliable. This unreliability can, however, be somewhat alleviated by combining scores.

Second, in terms of what is being measured, the rationale for the present independent weighting of scores on the orals is not clear: If the oral examinations are designed to measure content or knowledge then the written test also provides valuable information which is being disregarded in arriving at a score on each topic (e.g. Adult Orthopaedics); alternatively, if the oral examinations are designed to measure some specific ability related to oral confrontations then scores on all the oral tests should be taken into account simultaneously.

Third, the present practice with regard to the oral examination treats as equivalent, scores which most qualified observers do not regard as being of equal importance. For example, failing scores in Adult Orthopaedics and in Anatomy may or may not be of equal significance for the competent practice or orthopaedics. In contrast, present procedures necessarily treat these two failures as equivalent.

C. Setting Standards

1. The Written Examination

As noted above different techniques are employed in setting the pass-fail mark on the written and on the oral examinations.

In the written examination, the scores of repeaters and of graduates of foreign medical schools are removed from the distribution, the remaining scores are ranked and all scores one or more standard deviations below the mean are listed as failing. It follows that every year approximately 16% of the candidates taking the examination for the first time will necessarily fail.

This procedure has two advantages:

First, it eliminates the unintended fluctuation in standards that occurs when any fixed percentage (e.g. 75%) of correct responses is required to pass an examination. Experience has demonstrated that the difficulty of an examination can fluctuate markedly from year to year; consequently defining the passing level in terms of an arbitrary percentage of the examination means that a failing performance in one year may be a passing performance in another. Furthermore, there is no reason to believe that a fixed percentage of correct responses on a series of examinations has any relation to professional competence *per se*.

Second, assuming that the candidate population and the content sampled by the examination remain stable from year to year, the present procedure would yield a relatively stable standard, i.e., the same quality of performance would be passing from year to year. Further, if, after failing one year, a candidate actually improved substantially in whatever ability the examination measures, he would pass the next time around.

Both these advantages probably obtain in the short run but it is very dubious that they apply in the long run since both the content of the examination and the candidate population certainly change substantially over a period of time. In addition it should be noted that the present techniques for setting standards on the written examination have certain inherent weaknesses:

First, success on a certifying examination is supposed to indicate that the candidate has attained a certain level of professional competence at a certain point in time and failure is supposed to indicate that the candidate has some deficiency that impairs his professional performance. Unfortunately, at the present time there are no data to support the assumption that those (and only those) ranking in the bottom 16% on the written examination are necessarily unable to meet the standards of performance required of an orthopaedic surgeon.

Secondly, changes in the aspects of competence measured by examination or in the candidate population could quickly produce unintended alterations in the standards. For example, the increased candidate population in 1968 may be different in kind from that of previous years, or the war in Viet Nam may lead to an expansion of the number of exceptionally competent physicians deciding to specialize in orthopaedics.

Third, using a cut-off based purely on a "curve" can mask serious weaknesses, or obscure substantial improvements in training programs, since with the present procedures no matter how "poor" (or how "good") the training programs may be, the same percentage of the candidates will fail every year.

2. The Oral Examination

The oral examinations are treated as independent entities which must be passed or failed individually at an arbitrarily selected level (75%). The strength of the present system lies in the fact that individual examiners are able to perceive a direct relation between what the candidate says and the standards of competence required to perform specific tasks adequately. Thus, for example, the examiners in Adult Orthopaedics can assess a candidate's replies in terms of what they (the examiners) know is required of a practicing orthopaedist. The accuracy with which individual examiners in Adult Orthopaedics identified the fourth year men on the recent In-Training Examination supports this view. (See Document II.)

The weaknesses of this procedure, however, seem to outweigh its one strength:

First, there is a great deal of evidence that different examiners use different standards. The variation in the mean scores assigned by different examining teams emphasizes this fact. (See Document I)

Second, it is difficult to relate the content of some of the orals to on-the-job competence thus vitiating the great advantage of the orals. No one, for example, is quite sure how much basic science one needs to know to be an orthopaedist.

Third, the oral examinations are much less uniform, objective, and reliable than the written yet each one is treated as if it equaled the written test in importance and reliability. Some recognition of the unreliability of the scores on the orals is inherent in the provision allowing a candidate who fails only one to repeat that one, while requiring candidates who fail more than one to repeat the entire battery. However, that does not solve the basic problem.

IV. A Proposal for a New Procedure

The discussion of the strengths and weaknesses of present procedures implies some directions for improving practice. Specifically, any new system should have the following characteristics:

1. It should yield scores which directly relate to important areas of competence in terms of processes or content or both. The suggestions that have emerged from the work of the Examination Committee in developing the blueprint and that of the various Task Forces indicate that the following process scores should be considered:

    1. Ability to recall information
    2. Ability to interpret and analyze data
    3. Ability to solve problems
    4. Ability to relate to patients

2. Decisions regarding the combining process should take into account the reliability of the separate scores obtained, and the weights employed in making these combinations should reflect the relative importance of the various components of competence required for adequate performance in the practice of orthopaedic surgery.

3. The standards should be based directly on the critical performance requirements of orthopaedic surgery rather than on an arbitrary percent of an examination or of a candidate pool; they should be systematized in a fashion which ensures that the same level of competence is regarded as passing from year to year and should be changed only by explicit decision, not by fortuitous changes in the candidate population or the content of the examination; they should be readily understandable, easy to apply and recognized as appropriate by training chiefs and candidates; finally they should have a salutory effect on training programs.

Figure 1 is an example of a system that would meet these criteria. In this illustrative system the following characteristics should be noted:

1. Each candidate would get five scores: (1) Recall-clinical; (2) Recall-Basic Science; (3) Observation and Interpretation; (4) Problem Solving; (5) Ability to Relate to Patients. The recall score would be obtained from the multiple choice test only, while scores on Observation and Interpretation and on Problem Solving would be obtained by combining sub-scores on some parts of several tests (for example, the oral, the multiple choice, and the patient management problems). The score on ability to relate to patients would, at first, be obtained exclusively from the oral, but might ultimately include ratings from training chiefs.

2. All scores would be converted to a standard scale--in this example, a 12 point scale.

3. For each aspect of competence (i.e., each combination of scores) predetermined levels of "clearly failing" and "clearly passing" performance would be established, with the space in between representing a "zone of reasonable doubt." Candidates scoring above the "clearly passing" line would be certified; those scoring below the "clearly failing" line on any category would be denied certification regardless of how well they did with respect to other components of competence and they might be required to repeat part of all of the examinations; those with some scores falling in the "zone of reasonable doubt" might or might not be certified or might be given temporary certification with a deficiency and required to repeat certain portions of the examination within a given period depending on the specific guide lines developed by the Board. In addition, some appropriately weighted grand total score could be obtained and corresponding standards developed for it to

assure that no candidate was certified whose performance was marginal on all categories.

A major advantage of the system consists in the fact that it separates the act of measuring competence from the act of judging what to do about different patterns of competence and incompetence. That is, it provides the Board with a maximum amount of relevant and reliable data about a candidate on the basis of which a rational judgment can be subsequently be made in line with clearly specified "ground rules." To illustrate, three profiles representing quite different patterns of competence are recorded in Figure 1.

Candidate A's scores are above the clearly passing level in every category except "Ability to Relate to Patients." His score in that area is in the poor range; he might or might not be certified (or might be given temporary certification with a deficiency) depending on the ground rules established by the Board. Candidate B performed well on all the categories except Problem Solving. He would be denied certification because his Problem Solving score was below the clearly failing line; since his other scores were all well above the clearly passing level he might be required to repeat only some parts of the examination. Condidate C's scores were in the "zone of reasonable doubt" on all categories except on "Recall" where his performance is marginal; his grand total score would in all probability be fairly low; decision about his certification would again depend on the "ground rules" developed for handling marginal performance.

V. Steps in Implementing a New Procedure

Adoption of reporting procedures such as that illustrated above would be greatly facilitated by research designed to provide data to assist in determining the optimal weights and st standards employed. In obtaining these data the following next steps are recommended:

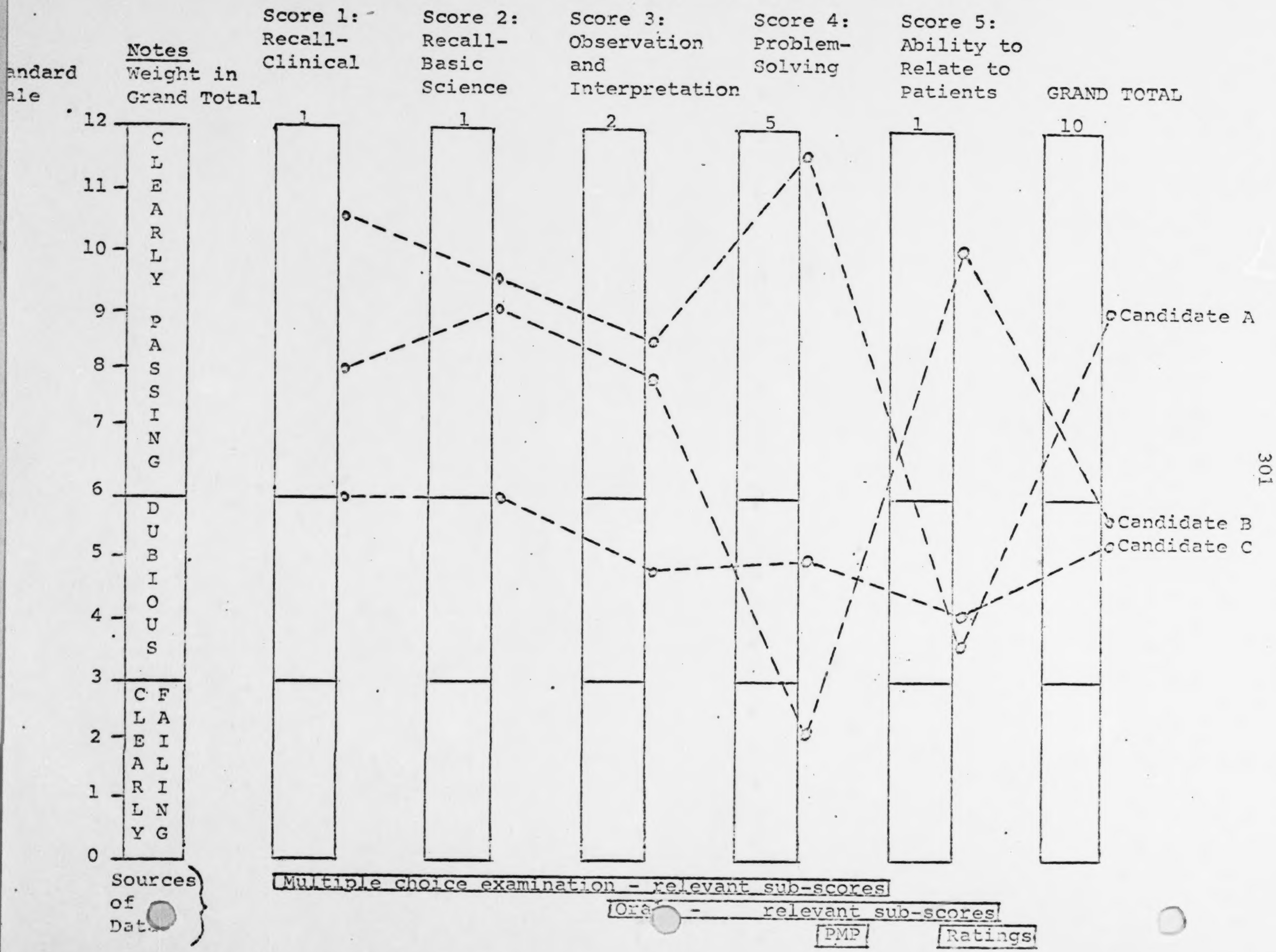
1. Review of the test materials by a committee to determine minimum passing levels for each question. (See Document III.) This technique is used in setting standards for undergraduates at the University of Illinois Medical School.

2. Further study of the validity of the various assessment techniques (a) by investigating certain hypothesized relationships among test scores (construct validity) and (b) by correlating test scores with the ratings of residents by qualified observers (concurrent validity). In the latter studies the scores of candidates rated as "poor" and the scores of those

beginning training can be considered in establishing examination standards related to performance per se.

Finally, on the basis of such data, together with other considerations outlined above, this Task Force should be prepared to make specific recommendations to the Board.

FIGURE 1: PROFILE OF PERFORMANCE



DOCUMENT I

January 1966 Orthopaedic Certification Examination

Mean Scores of teams of examiners on conventional oral examinations and the mean multiple choice score of the candidates evaluated by each team

Panel	N	Pathology	Childrens	Anatomy - Trauma	Adult	Multiple Choice (Raw)
A	29	86.3	83.1	86.7	82.5	109.3
B	29	79.9	84.0	82.7	81.9	103.0
C	29	79.0	84.0	83.0	80.7	106.0
D	29	77.8	82.6	80.9	79.7	102.3
E	29	82.2	86.9	82.5	81.8	107.5
F	30	82.8	84.6	75.4	77.8	102.7
G	28	77.5	77.3	84.0	77.9	101.1
H	30	80.7	80.6	80.8	82.2	105.2
I	29	80.8	83.0	86.2	82.0	107.0
J	30	82.3	79.4	77.3	83.6	103.8
K	28	86.1	80.8	81.6	84.4	105.2
L	28	78.6	81.3	82.6	81.1	107.4
M	28	78.1	84.5	83.3	81.8	105.2
N	7	80.9	80.3	80.7	77.9	102.1
TOTAL	383	81.0	82.5	82.0	81.3	105.0

DOCUMENT II

MEAN OVERALL SCORES ON IN-TRAINING ORALS BY YEAR OF GRADUATION AND YEAR OF TRAINING

Year of Training		Year of Graduation							
		1965	1964	1963	1962	1961	1960	1959 and earlier	Total
1	N	14	6	5	0	1	0	3	29
	DI	5.4	6.2	4.0		10.0		4.0	5.4
	PT	6.3	6.1	5.4		11.0		5.5	6.2
	AO	65.6	58.7	66.0		70.0		71.0	65.0
2	N	0	33	6	19	8	5	4	75
	DI		7.0	6.3	7.4	6.1	6.2	4.3	6.8
	PT		7.0	5.2	7.1	7.1	8.6	5.3	6.9
	AO		69.8	61.3	71.4	70.6	73.8	76.3	70.2
3	N	0	0	18	11	10	1	10	50
	DI			7.3	6.5	7.1	3.0	7.0	6.9
	PT			6.0	7.6	6.6	3.0	6.5	6.5
	AO			77.4	71.2	77.6	90.0	68.9	[illegible]4.6
4	N	0	0	0	25	15	20	19	79
	DI				7.9	6.9	8.0	7.2	7.6
	PT				7.4	7.6	8.0	7.0	7.5
	AO				81.0	80.9	81.5	74.3	79.5
Total		14	39	29	55	34	26	36	233
	DI	5.4	6.9	6.5	7.4	6.9	7.5	6.6	6.9
	PT	6.3	6.8	5.7	7.3	7.3	7.9	6.6	6.9
	AO	65.6	68.1	72.1	75.6	77.2	80.4	72.8	73.6

N = Number of Residents
DI = Diagnostic Interview
PT = Proposed Treatment Interview
AO = Adult Orthopaedic Examination

All examinations were administered to Residents at all levels of training by one examiner on the two Interviews and one for Adult Orthopaedics. The examiners did not know the residents' level of training.

Appendix 26

SETTING STANDARDS OF COMPETENCE

THE MINIMUM PASS LEVEL

Prepared by

Committee on Student Appraisal

University of Illinois College of Medicine
January, 1964

I. Theoretical Considerations

The Committee on Student Appraisal has explored alternative methods by which the faculty can evaluate student competence as precisely and as validly as possible. Basically, there are only two ways of assessing an individual's performance on any achievement test, including our own comprehensives and departmental examinations: (A) by judging his performance in terms of its relation to others in his group; and (B) by judging performance on the basis of pre-determined criteria.

A. "Grading on the curve" (i.e., judging by relative standards) is widely used throughout the educational system, including with such standardized tests as MCAT and National Boards. However, even under optimal circumstances the method is subject to the following deficiencies:

1. Standards are set in terms of what is rather than in terms of what ought to be.
2. An individual will "look good" either if he personally achieves a great deal, or if he is a member of a group most of whom achieve very little. It follows that some students who are certified as satisfactory in one year would not necessarily have been so certified had they belonged to the group a year earlier or a year later.
3. Changes in the quality of group performance cannot be accurately observed, documented and thus, manipulated, since performance on each test is measured in relative terms that provide inadequate basis for comparison with performance on other tests.

These deficiencies are especially serious if the reference group is a single class, in a single year, in a single school, from a highly select population (such as nuclear physicists or medical students). In this situation the group itself becomes the arbiter of the standards in terms of which it is to be judged and group enforced pressures can be as serious in schools as in factories.

B. Alternatively judging the individual's performance in terms of criteria of adequacy which are independent of the performance of the particular group of which he is a member, may have the following drawbacks:

1. Responsible faculty members do have different standards.
2. Even if agreement on general standards can be reached, there remains the difficulty of obtaining specific agreement about what level of performance on a particular test is necessary to meet that generally accepted standard.
3. Experience with the specific type of test is required in order to determine exactly what is being tested.
4. Finally, revisions in standards may be required in order to bring them into accord with the reality of a particular learning situation. (For example, it may be desirable, but "unrealistic" to require that all medical students develop facility with clinical research techniques if that goal is assigned a lower priority than clinical competence per se, and if time to accomplish both is inadequate.)

Despite these inherent difficulties this method is uniquely suited to the evaluation of individuals who come from a highly select population and who belong to a single community of scholars who share a common concept, however vaguely defined, of what constitutes acceptable professional performance. It has the specific advantage that the faculty does, in fact, establish the standards for student performance. The application of this method to comprehensive examinations requires (a) that the standard be set in terms of a level of competence, not in terms of relative position in a group, and (b) that this level of competence reflect an explicit judgment about the character of individual questions and problems, not merely an overall estimate of what a student "ought to do on tests like this."

II. Procedure

After careful consideration of the feasibility and possible usefulness of a pre-determined minimum passing level for our internal, certifying examinations, the Committee on Student Appraisal has reached the following conclusions:

1. Definition - The minimum passing level (MPL) is the minimum number of questions a student must answer correctly in order to be certified as having done satisfactory work in a course. It is of necessity based on a large number of items.

2. It is recognized that such an MPL score is based largely on the subjective judgment of one or more faculty members and that it may vary appreciably among faculty members.

3. The use of the MPL provides a concise opinion as to the relative difficulty of the individual items and of the test as a whole. Also, it will provide an accurate means of comparing different portions of a total examination, and different examinations.

In view of these considerations the following procedure is recommended to the departments for establishing the recommended passing score on the multiple choice questions in the comprehensive examinations:

> Consider each question separately; review carefully the total list of choices given in answer to each and decide which ones a student "who knows enough to pass the course" should be able to reject. For example, if a question has 5 possible answers given, only one of which is correct, and it is judged that the "student who knows enough to pass the course" should be able to reject one of the wrong choices, it follows that the barely passing student could be expected to answer such a question correctly approximately one-fourth of the time. If the test contained a large number of such questions the barely passing student could be expected to answer correctly at least 25% of them. This is the first step in establishing a minimum passing level. This initial estimate for an individual item should then be revised by a small amount depending

on the degree of discrimination required, in the opinion of the estimator, to select among the 4 remaining answers. This final estimate is to be recorded as the MPL for the individual item. The estimates for each item are then averaged and the resulting figure recorded as the overall Minimum Passing Level for that portion of the test.

EXAMPLE I

Light has wave characteristics. Which of the following is the best experimental evidence of this statement?

A. Light can be reflected by a mirror.

B. Light forms dark and light bands on passing through a small opening.

1/4
MPL = .25*

C. A beam of white light can be broken into its component colors by a prism.

D. Light carries energy.

E. Light operates a photoelectric cell.

* Staff decided no upward revision of the MPL was required.

In deciding on the MPL it is important to note that a question about a very sophisticated concept may require only the most elementary understanding if the wrong answers are so implausible as to require very little discrimination in rejecting them, as in Example 2; alternatively, the range of choice given in answers to the identical question may require very fine discriminations as in Example 3.

EXAMPLE II

Which of the following is the best estimate of the current population of the United States?

1/1
MPL = 1.00

A. 2,000
B. 200,000
C. 2,000,000
D. 20,000,000
E. 200,000,000

EXAMPLE III

Which of the following is the best estimate of the current population of the United States?

~~A.~~ 176,000,000
1/4 B. 178,000,000
MPL - .25 C. 180,000,000
D. 182,000,000
E. 184,000,000
~~F.~~ 186,000,000

Clearly the percent of such questions that the knowledgeable student should answer correctly differs markedly in examples 2 and 3. For this reason it is essential to determine the minimum passing level on the basis of careful inspection of each response to each item, not on the basis of some general, overall estimate of the test as a whole.*

The MPL is obviously of little significance unless it is based on a thorough analysis of every item in a test including consideration of incorrect alternatives as well as correct responses. The validity of the estimate of the MPL also depends on obtaining independent judgments from several people in the department who are familiar with the nature and objectives of instruction at the level for which the test is designed. The usefulness of the estimate will be greater, the greater the number of knowledgeable faculty examining the test. In departments containing a number of sub-specialists who are not in close contact with subject matter taught by their colleagues, not all staff members will be in a position to make informed judgments about the MPL. Ultimately, however, it is desirable that standards be developed on the basis of wide discussion by all or most members of the teaching staff of the department.

---

* For certain multiple choice items and objective questions in other formats the recommended procedure may be difficult to apply. In such cases the following method should be employed:

> Consider each question separately, review carefully the total list of choices given, considering especially the subtlety of the discrimination required. For each item determine what percent of such questions the barely passing student should answer correctly if the test were to contain a very large number of the same type of questions, about the same subject. Record this percent as the MPL for that item. The estimates for each item will then be averaged and the resulting figure recorded as the overall Minimum Passing Level for that section of the test.

If differences among the several judgments are relatively small the extremes can be used to define a "zone of doubt" below which performance is considered clearly failing and above which it is considered clearly satisfactory. For example, if the average of one instructor's estimate of the MPL is 43% and that of another is 45% and that of a third is 47% - the department may recommend that scores below 43% are clearly failing; those above 47% clearly passing; and those ranging from 43% to 47% inclusive are within a "zone of doubt" to be interpreted in light of the student's performance in other disciplines. However, if the discrepancies among the individual estimates are large, clarification of departmental standards and grades could serve as a basis for reconciling differences.

III. Experience with Pre-determined Standards

Information about the use of such standards has, until recently, come primarily from non-educational situations. For example, in the definition of job specifications and the testing of a particular applicant group, the percent who "pass" or "fail" does not affect the standard for passing or failing. For example, among applicants for a position as clerk-typist at the University of Illinois College of Medicine, only those who are able to type 60 words a minute can be "certified" as meeting acceptable standards. In the determination of eligibility for a license to drive a car it is required that the applicant know the shape of all common road signs indicating potential danger. In addition to personnel and licensing agencies some liberal arts colleges, notably the University of Chicago, have used the technique.

Until approximately 1946 grades on the introductory course in Social Sciences at the University of Chicago were determined exclusively on the basis of a "curve". A distribution of student scores was made and the distribution inspected to find "breaking points" that would yield approximately 10% F's and 15% A's. The letter grades B, C and D were determined in a similar fashion though with somewhat less rigid adherence to a pre-determined frequency. In the Fall of 1946 there began to be numerous comments among the staff that "it is such a pleasure to teach this group of students," "they seem so much better" or "they are so mature". The striking change in the quality of the student body was attributed by the staff to three related factors: Among entering college students there was a high proportion of veterans who for a number of years had been in a position to observe (consciously or not) the responses to stress situations of individuals, groups and institutions in widely varying cultural contents. These students were thought to be more "mature" in that they had a greater first hand knowledge of social and social-psychological phenomena. Secondly, the very existence of the veteran group expanded the applicant pool from which selection was made. Third, the more "mature" individuals were thought to be "more serious" in their approach to study and this in turn was credited as affecting the climate for learning of the highly selected younger students. All these influences did indeed make plausible the view that "this was a much better class". In view of the extensive reference during the year to this phenomenon, the Examiner in Social Sciences asked the staff to reconsider its procedures for assigning grades on the year-end comprehensives. The procedure decided upon was that closely similar to the one described above as being advised for liaison examiners in this College.

It may be of interest to note that over a period of approximately 12 years the percent required as a minimum passing score varied within quite narrow limits (2% to 4%), until a major innovation in the nature of the examination questions was introduced. In the last few years prior to the employment of a pre-determined standard, the failure rate had been 10%-12%; in the first few years after its employment the failure rate fell to 2% or 3%; when the group of veterans moved out of the undergraduate program and the applicant pool was further reduced by general population trends (fewer college age students 20 years after the depression) the failure rate rose again for a brief period to an all time high of 16% to 18%. Independent, concurrent studies of the performance of the several entering classes on admissions and placement tests led to a predicted variation in the failure rate which coincided closely with the actual rate obtained by using the pre-established MPL.

A similar method was employed in the Physical Sciences at the University of Chicago. The procedure and the results obtained over a two-year period are reported in detail by Dr. Leo Nedelsky*, who concluded that within the limits of that pilot study the following questions were answered "in the affirmative, with varying degrees of conclusiveness:"

---

* Nedelsky, Leo, "Absolute Grading Standards for Objective Tests". Educational and Psychological Measurement Vol. 14, Spring, 1954. pp 3-19.

"1. Can a group of instructors teaching the same course agree on the minimum passing score using this technique?

"2. If question 1 is answered in the affirmative, in the sense that the instructors independently arrive at nearly the same score, do they agree on the relative difficulty of individual test items? If they do, the agreement on question 1 is detailed rather than merely statistical and thus has greater theoretical meaning.

"3. If question 2 is answered in the affirmative, do the instructors agree in their selection of (unsatisfactory) responses? If they do, the agreement in 2, and consequently agreement in 1, acquire greater theoretical meaning.

"4. Is the instructors' judgment of the difficulty of individual questions sound; that is, do the students' scores support it?

"5. Is the instructors' identification of (failing) responses sound, that is, do the students, or at least the non-failing students, find these responses less attractive than other wrong responses?

"6. Is the basic assumption of the theory underlying the proposed technique sound; that is, are there identifiable responses which play an important role in differentiating between failing and non-failing students?...If the proposed technique is to be theoretically sound and not a mere computational device, the failing and the non-failing students should show a measurably different reaction to the clearly unsatisfactory responses.

"7. Does the application of the technique to different examinations result in minimum passing scores corresponding to the same standard of achievement?"

The principle of setting standards in advance of the administration of the examination was first employed at the University of Illinois College of Medicine with the experimental Sophomore Comprehensive in June, 1962. Those very conservative estimates (25%-30% correct answers) identified 13 failing grades made by 12 students. For these 12 students the following have been inspected: the 2 year G.P.A., the National Board average and the scores on the Junior Comprehensive. All but one of the students are at a satisfactory level on these measures (albeit, in some cases, a barely satisfactory level).

With the exception of the performance on the Junior Comprehensive, such results are not surprising in view of the fact that the comprehensive was deliberately designed to measure a type of competence which it was believed had been underemphasized in the other measures of achievement. Had the results on the comprehensive identified as failing the same students as were so identified by other methods of appraisal, the comprehensives would have contributed very little to the certification process, per se. As regards the relation between performance on the Sophomore Comprehensive of 1962 and the Junior Comprehensive of 1963, substantial change in the class standing of individual students often occurs between the basic science years and the clinical years. As we succeed in utilizing the comprehensives to test the basic understandings common to the entire medical program such shifts could be expected to diminish.

In 1963 there was an attempt to utilize the concept of a predetermined standard in setting standards for all four comprehensives. Except for the Freshman Comprehensive the recommended standards resulted in a failure rate which appeared excessive by any criterion. Whether this was due to some combination of inadequacies in the examination, the students or the teaching, or whether it was due to misunderstanding or inexperience with the method of setting the MPL cannot be determined on the basis of the data now available. However, even without additional data it is important to note the discrepancy between expectations and performance and to investigate ways of reducing it. The least discrepancy occurred on the Freshman Comprehensive; it is felt that this was due in large part to the fact that liaison examiners in the Freshman year had (by the end of the year) employed the method on three occasions and were therefore able to develop standards in which they could place greater confidence. Experience with the Freshman Comprehensive of December, 1963 supports this interpretation in that application of the pre-established MPL yielded results that were consistent with other evidence about the class, as interpreted in light of previous experience.

In the final analysis the most significant issue is whether the comprehensive examination system distributes students appropriately and whether it encourages them to develop the essential kinds of competence.

Appendix 27

Working Papers for: the Meeting of
TASK FORCE CHAIRMEN and EXAMINATION COMMITTEE

INTRODUCTION

In the course of the Orthopaedic Training Study numberous more or less independent recommendations have been made to the American Board of Orthopaedic Surgery regarding its certification procedures. Some of these have already been implemented; others await action by the Examination Committee and the Board. It is the purpose of this meeting to review each of the recommendations now available from the several Task Forces and to prepare a unified set of recommendations regarding the entire certifying process, for action by the Board. In order to give perspective to this task, the recommendations that have already been implemented are listed first, followed by those that require further action by the Examination Committee. Relevant supporting documents are included as Appendices.

RECOMMENDATIONS ALREADY IMPLEMENTED

Recommendation 1: The Board shall establish an examination blueprint specifying the subject-matter content and the cognitive processes to be evaluated and the weight to be assigned each in the certifying examination. This set of specifications shall govern the overall make-up of the examination. (See Appendix A).

Recommendation 2: The Examination Committee shall establish regular procedures for maintaining and updating the calssification of materials in the item pool in accord with the categories in the blueprint. (See Appendix B).

Recommendation 3: The Board shall establish task forces for the purpose of developing, reviewing and refining new materials for use in the multiple choice and patient management components of the examination. (See Appendix C).

Recommendation 4: The following procedure shall be followed in the preparation of the written examination: The technical staff will draw from the pool of new and old materials, questions and case materials to meet the specifications of the blueprint, but in excess of the number required for a single examination. These materials will be circulated to the Examination Committee in EXAMINATION FORMAT and without a key to the correct answers, for independent, individual response by each Committee member. An item analysis of these responses will be prepared, and selection of questions to be included in the written examination will be determined by these findings. A final copy of the examination will then be circulated for review and action by the Examination Committee, at least three months before the examination date.

RECOMMENDATIONS REQUIRING ACTION

Recommendation 5: The total certifying process shall be designed to yield evidence on the following aspects of professional competence:

I Surgical skill
II Professional habits and attitudes
III Ability to recall information
IV Ability to interpret and analyze data
V Ability to solve problems (including clinical judgment)
VI Ability to relate effectively to patients and colleagues

As soon as feasible provision will be made to obtain a separate score on clinical judgment.

Recommendation 6: Evidence on Factors I and II will be gathered primarily through questionnaire and rating scales, to be completed by program supervisors and evaluated by the Committee on Eligibility. Admission to the certifying examination will be granted only to those candidates who meet satisfactory standards on these dimensions of performance. (See Appendices D through G).

Recommendation 7: Evidence on Factors III through VI will be gathered primarily through written and oral examinations as follows:

A - The Multiple Choice Examination will be designed to yield on Factors III, IV, and V.
B - Patient Management Problems will be utilized as a regular part of the certifying examination to yield evidence on Factor V (and, where relevant, on Factor IV). (See Appendix M).
C - The Oral Examination will be redesigned to yield evidence on Factors IV, V and VI and limited evidence on Factor III. (See Appendices D and H). To accomplish this goal it is specifically recommended:
    (1) that the oral examination be made up of: (a) three half-hour examinations designed to assess problem-solving skills, ecch of which to be administered by one examiner utilizing previously prepared case materials; (b) one half-hour examination designed to assess interpretive skill to be administered by one examiner utilizing previously prepared X-ray and similar material. (See Appendix I for possible revision of this recommendation); (c) one half-hour examination designed to assess skill in relating to patients in simulated physician-patient encounters, to be administered by two examiners utilizing standardized case materials.

(2) that each component of the oral examination be scored on specially prepared rating forms analogous to that used in the Simulated Patient Interviews. (See Appendi J and K).

NOTE: Acceptance of this recommendation requires development of procedures for preparation of more nearly standardized oral examination materials and for examiner training.

Recommendation 8: The present system of treating each examination independently shall be replaced by a method designed to describe a "profile" of performance and shall be reported in relation to predetermined standards. (For a description of the rationale behind this recommendation see Appendix L). Specifically, it is recommended.

A - that each candidate's scores be reported as shown in Figure 1. (For details on the mechanics of obtaining these scores see Appendix N).

NOTE: If for reasons of convenience, it is preferred to report candidate performance in tabular rather than graphic form the profiles shown in Figure 1 can be reported in the manner illustrated in Table I following.

B - that the following ground rules be established:

(1) that any candidate who falls within the "clearly failing" range on the total OR an ANY sub-score III through V (i.e., recall, observation and interpretation, or problem-solving) be required to repeat the entire examination; (e.g., Candidates B and C Figure 1).

NOTE: Acceptance of this ground rule would have the effect of requiring some individuals who have a "clearly passing" score on the Grand Total to repeat the examination. For example, Candidate C, Figure 1, has such high scores on Factors III, V and VI that he achieves a "clearly failing" score on the Grand Total despite his "clearly failing" score on Factor IV. However, requiring him to repeat the examination does not differ in principle from the ground rules regarding eligibility that have the effect of excluding some applicants who would undoubtedly pass the certifying examination.

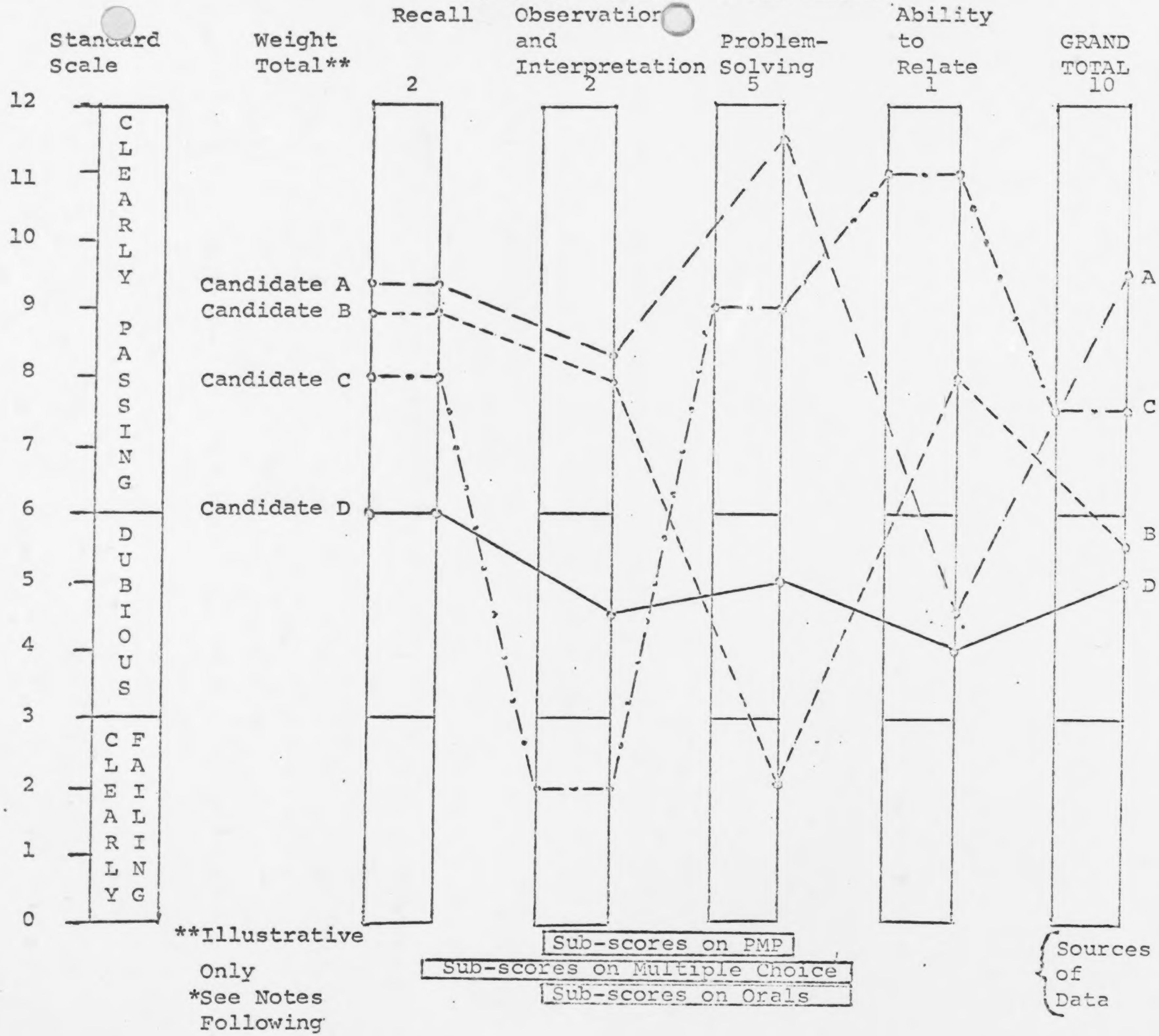
(2) that the total profile be reviewed for any candidate who falls within the "dubious" range on the total OR in the "clearly failing" range on sub-score VI (i.e., ability to relate) and ground rules be established for disposition of such cases; (e.g., Candidate D, Figure 1).

(3) that all other candidates (i.e., those whose scores a are above the "clearly failing" level on all subscores, and within the "clearly passing" range on the total) be certified; (e.g., Candidate A, Figure 1).

(4) that until sufficient experience is gained with this system the Board reserve the right to modify the pre-established standards in the event they result in unacceptable failure rates. (For techniques of establishing pre-determined standards see Appendix L, Document III and pp. 8-13 in Appendix M).

NOTE: Acceptance of Recommendation 8 requires a change in the time at which the Board has traditionally announced the results of the examination.

FIGURE I: PROFILE OF PERFORMANCE*



**Illustrative Only

*See Notes Following

NOTES ON FIGURE 1

1 For a description of the teachnque of obtaining these scores see Appendix N.

2 The weights assigned to each major score (Recall, etc.) should be a matter of general policy and could be expected to remain relatively constant from year to year.

3 The relative contribution of each test to each score (e.g., multiple choice and oral in the recall score) would need to be determined each year in light of the specific characteristics of that examination. (See Appendix N).

4 It is expected that a separate score on Clinical Judgment will be added to this profile as soon as feasible.

Table I: Alternative Format for Reporting Individual Profiles

	Factor III Recall	Factor IV Observation and Interpretation	Factor V Problem-Solving	Factor VI Ability to Relate	Grand Total
	Weight 2	Weight 2	Weight 5	Weight 1	
Candidate No.					
A	9.3	8.4	11.5	4.6	9.8
B	9.0	8.0	2.0	8.0	5.2
C	8.0	2.0	9.0	11.0	7.6
D	6.0	4.4	5.0	4.0	5.0

Appendix 28

Process Analysis of
Oral Patient Management Problems, Interpretive Skills and Role-Playing Examinations

by Christine McGuire

In January, 1968 a cadre of approximately 200 trained examiners, utilizing identical sets of standardized case materials, administered 5 half-hour oral examinations to each of 854 candidates for certification by the American Board of Orthopedic Surgery. For each candidate one half-hour consisted in role-playing simulations devoted primarily to assessment of skills in relating to patients and colleagues (I), one to analysis of x-rays, slides and other visual materials chosen to assess observations and interpretive skills (II), and the remaining 3 to diagnosis and recommendations for therapy in cases selected for assessment of problem solving skills in three clinical disciplines (III). In order to estimate the extent to which each type of examination sampled the behavior it was designed to elicit, a systematic observational analyses was made of a stratified random sample of the over 4,000 individual examinations administered to the 1968 candidate population.

Method

The sample of examinations to be observed was drawn and assigned to observers so as to be representative of the entire population with respect to the following variables: type of examination (i.e., I-simulation, II-interpretation, III-problem solving), content (i.e., adult orthopedics, children's orthopedics, trauma), examiners, candidate and time of administration (i.e., early morning, late morning and afternoon of each day). To assure appropriate representation of each variable, the observations were arranged so as to obtain a minimum of 40 of each type of examination and of each content area, at least 1 of each examiner, a random sampling of candidates with no more than 1 observation of a given candidate, and at least 35 observations at each major time period. Assignments were allocated to each of the 11 observers in a manner to assure that the observations of each member of the team were appropriately distributed with respect to each variable and to obtain a set of duplicate observations in which each observer was paired at least once with a physician member of the team and at least once with a specialist in educational evaluation.

In the actual observations only a few departures from the planned sampling occurred as a consequence of some candidate absenteeism or of observer delay in locating the room of an assigned observation. As finally accomplished a total of 235 observations (153 single and 41 duplicate observations) were made of 194 individual half-hour examinations.

The observer team was composed of 6 physicians, 2 trainees in educational research, and 3 specialists in educational evaluation.

Utilizing video tapes of sample examinations, this team was trained in the application of a modified form of interaction analysis to the descriptive recording of examiner and examinee behavior in the three types of oral examinations to be observed. The observers were instructed to record each unit of verbal behavior by utilizing a set of designated symbols to identify the characteristics, sequence and initiator of each exchange during the half-hour period. The categories that were employed in describing the behavior of each party are shown in Table I.

TABLE I: BEHAVIORAL CATEGORIES

Types of Examiner Behavior	Types of Candidate Behavior
1 Asks for information	1 Requests general information or clarification
2 Asks for specific interpretation, conclusion or recommended action	2 Requests specific data
3 Asks for reasons, evidence or criteria	3 Gives information, generalization (recall)
4 Gives general instruction or information	4 Gives specific observation or interpretation
5 Gives specific data	5 Gives intervention, inference, summary conclusions
6 Gives clarification or interpretation	6 Gives principle or reason on demand
7 Provides cues	7 Gives evidence of empirical validation on demand
8 Expresses hostility or challenge	8 Expresses hostility, rejection
9 Asks for reassurance or understanding	9 Persuades, influences, manipulates, reassures

NOTE: The meanings of Category 1 of examiner behavior and Category 3 of candidate behavior differ in the simulation (I) and non-simulation (II and III) examinations. In the former, an examiner, playing the role of a patient may request information about "his" illness (to be coded 1 or 9 depending on the predominant element in the request), and the candidate may respond with specific information (coded 3) or with reassurance (coded 9) or may ignore the plea for help (coded 8). In the non-simulation examinations these symbols take on a conventional meaning applicable to any type of oral examination.

Unlike other forms of interaction analysis, in the variant employed in this study the recording unit was behavior, not time. Observers were therefore instructed to make an entry when, and only when, either the speaker or the nature of his verbal behavior changed. Thus, a single symbol might in one case represent a piece of verbal behavior of only a few seconds' duration, and in another case the same type of verbal behavior of several minutes' duration.

In the objectified form developed for recording the observations provision was also made for the observer to indicate the initiation and termination of each exchange, the point at which the general context (topic, problems, etc.) of the discussion changed, the time at which this occurred and the nature of the stimulus producing it.

Findings

Table II summarizes the findings regarding the number and nature of examiner-candidate exchanges in the three types of orals analyzed. The data indicate that in the role-playing simulations of physician-patient and physician-colleague encounters (Type I), the behavior of the examiner (who took the role of patient or colleague) was most frequently 22% of all entries = 44% of examiner units characterized as an expression of hostility or challenge or a plea for reassurance and understanding. The second most common type of examiner behavior (16% of all entries = 32% of examiner units) consisted in requests for general information or specific interpretation, which in this type of oral were usually of the form, "Doc, how long will I be in the hospital?" or "Will I be able to walk without a cane?" i.e., questions of the type patients commonly ask that can be satisfied without calling on a large fund of esoteric information. In reaction to these challenges and inquiries, candidates tended to make cognitive responses (28% of all entries = 56% of candidate units) almost twice as frequently as affective responses (17% of all entries = 34% of candidate behavior). Among the affecting responses persuasion and reassurance were far more frequent than hostility and challenge (15% and 2%, respectively) though it is of interest that the latter occurred at all in an examination situation in which the candidate was making every attempt to respond in the most professionally acceptable mode.

In contrast with the findings for Type I orals, affective behavior on the part of either the candidate or the examiner was insignificant in both the interpretive (Type II) and problem-solving (Type III) orals.

TABLE II: NATURE OF EXAMINER-CANDIDATE BEHAVIOR

Type of Examination	Number of Observations	Percent of Total Behavioral Units Recorded As:																		Average No. of Interactions per half-hour	Average No. of discrete topics or problem-situations per half-hour
		Examiner Behavior Category:*									Candidate Behavior Category:*										
		1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9		
I-Simulation	53 (8 duplicate)	7	9	1	2	6	1	1	12	10	1	6	15	3	10	1	0	2	15	46	2.7
II-Observation and Interpretation	45 (9 duplicate)	6	27	3	3	5	2	4	1	0	1	3	8	19	15	2	1	0	0	49	2.7
III-Problem-Solving	137 (24 duplicate)	4	16	2	2	18	2	2	1	1	1	19	6	3	20	2	1	0	1	58	1.8

* For definition of Categories see Table I.

In the former, as might be expected, the predominant form of examiner behavior (54% of all examiner entries) was characterized as "asking for a specific interpretation, conclusion or recommended action;" approximately 70% of candidate behavior was a direct cognitive response to inquiries, specific observations and interpretations being slightly more numerous than general summaries and conclusions. Interactions in the problem-solving orals (Type III) were significantly different in that approximately 40% of the entries represented exchanges in which the candidate asked the examiner for additional specific data needed to make a diagnosis or to decide on the next steps in the management of a standardized case presented for solution.

It is of interest to note that there were also significant differences among the three types of examinations with respect to the number of exchanges and the nature of the behavior initiating each exchange. Approximately one-third more interactions were recorded for the problem-solving orals (Type III) than for the simulation (Type I). This was in part accounted for by the fact that in the former one-third of the exchanges were of a "rapid fire" character initiated by the candidate's request for specific historical, physical, laboratory or x-ray findings in the patient presented for his analysis. In contrast, in the simulation exercises (Type I) about 90% of the exchanges were initiated by the examiner employing the kinds of gambits patients commonly use to obtain information or reassurance. Also in the interpretive exercises (Type II) it was the examiner more often than not who initiated the exchange, usually with a question demanding a specific interpretation.

These findings are substantially different from those obtained in a similar observational analysis made in January, 1965 of the traditional orals which the Orthopedic Board administered at that time. * Two members of the current study team also participated in that survey. It was their general impression that in contrast with the examinations of 1965 which were conducted as oral quizzes designed to test the candidates ability to recall a vast body of factual information rapidly and under stress, the examinations of 1968 were conducted in a manner that gave the candidate time to think about and opportunity to pursue an approach to a problem. The data reported in Table III support this impression with respect to the nature, but not with respect to the number of candidate-examiner exchanges tallied in the two studies.

TABLE III: COMPARISON OF 1965 AND 1968 ORAL EXAMINATIONS

Date	No. of discrete topics or problem situation	Av. No. of exchanges per 1/2 hour	Percent of Candidate Behavior Involving Primarily:			
			Simple Recall	Interpretation of Data	Problem Solving	Other
1965	*	43	69	18	13	***
1968	2.2	54	15	11	59	15

* Not separately tallied though in most observations the topic shifted after every 2-3 exchanges and, in some instances shifted with every exchange; 12-15 topics per half-hour would constitute a conservative estimate.

** For 1968 the following categories of candidate behavior are included:

Recall = Candidate Category 3 - Gives Information, etc. Interpretation of Data = Candidate Category 4 - Gives specific observation or interpretation. Problem Solving = Candidate Category 2 - Requests specific data, plus Category 5 - Gives intervention, etc.

*** Not applicable in the 1965 study.

Due to certain artifacts in the recording systems employed in 1965 and 1968, the data presented in Table III probably underestimate the differences in the oral examinations at the two periods. However, even without any allowance for this conservatism, they provide strong support for the view that the pattern of competence assessed by the two types of examinations is significantly different.

Two types of evidence bearing on the reliability of the observations are also relevant in interpreting the data presented above. The first, reported in Table IV, indicates that there were significant differences among observers in the average number of behavioral units each identified and in the proportions of entries each recorded in the several behavioral categories. While these differences are in large

TABLE IV: RANGE OF FINDINGS ON EACH OBSERVER'S TOTAL OBSERVATIONAL SAMPLE*

Type of Examination	Value Recorded by any Observer	Average No. of Behavioral Units Recorded Each Half-hour	Percent of Behavioral Units Assigned to:																	
			Examiner Category**									Candidate Category**								
			1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
I-Simulations	Highest	136	16	15	2	6	11	4	3	18	20	3	11	23	6	19	8	1	8	28
	Lowest	86	1	0	0	0	0	0	0	2	4	0	0	2	0	0	0	0	0	6
I-Observations and Interpretation	Highest	129	12	35	11	9	8	4	6	4	0	3	8	15	35	22	8	2	0	1
	Lowest	77	0	16	1	0	2	0	2	0	0	0	0	1	8	8	1	0	0	0
III-Problem-Solving	Highest	153	9	20	3	4	24	5	4	3	1	4	23	11	4	20	6	2	0	2
	Lowest	101	2	11	1	0	15	0	0	1	0	0	14	3	2	14	0	0	0	0

* This table is to be read as follows: Considering each observer's sample of observations as a random sample of the total population, the range of findings in these sub-samples is given by the highest and lowest value recorded by any observer.

** For definition of categories see Table I.

measure attributable to real variation in the population of candidates and examiners assigned to each observer, Table V indicates that they cannot be wholly accounted for by this factor. These latter data suggest that at least a part of the inter-observer variance is error variance due to some discrepancies among observers not only in their application of the several categories but also in their interpretation of what constituted a "unit of behavior." These latter differences influence not only the number of interactions recorded but also the proportion of entries in each behavioral category, since such discrepancies most often occurred eight at points where there was a series of inquiries and responses of the same type on the same topics,* or at points where one of the parties to the interaction engaged in a relatively lengthy discussion that entailed movement back and forth from the behavioral category to another.

Conclusions

Given the inter-observer variation reported in Tables IV and V comparisons between candidates or between examiners are not warranted in view of the extremely limited observational sample of each. However, given the general similarity in the patterns of findings on the duplicate observations and the fact that the method of assigning observers was such as to randomize observer bias, the data appear to be sufficiently reliable to support two important conclusions:

(1) The pattern of competence sampled by the standardized oral examinations administered in 1968 differed significantly from that sampled by the unstandardized oral quizzes previously utilized; and

(2) In the three types of oral examinations administered in 1968 the element of competence most heavily weighted by each differed from the other two in the expected direction.

Subsequent factor analysis of the 1965 and 1968 written and oral examinations, together with a multivariate repression analysis of the concurrent validity of the 1968 oral examinations, has yielded additional evidence in support of these conclusions.

---

* For example the inquiries: "Hemaglobin? Hematocrit? CBC?" and responses to these inquiries were in some instances recorded as 3 exchanges (i.e., 6 behavioral units) and in others were recorded as 1 exchange (i.e., 2 behavioral units).

TABLE V: COMPARISON OF FINDINGS ON DUPLICATE OBSERVATIONS WHEN EACH MEMBER OF TEAM WAS PAIRED WITH A PHYSICIAN AND AN EDUCATOR

Type Examination	No. of Duplicate Observations	Observer	Average No. Recorded Each Half-Hour		Percent of Behavioral Units Assigned To:																	
			Inter-actions	Behavioral Units	Examiner Category:*									Candidate Category:*								
					1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
-Simulation	5	Physician with	51	124	5	14	1	3	5	5	1	7	3	3	6	12	3	18	1	0	1	11
		Other	48	119	6	7	1	8	2	0	1	13	11	1	7	16	3	6	3	0	4	12
	3	Educator with	37	95	5	0	2	0	1	0	0	21	13	0	1	19	0	0	0	0	1	34
		Other	43	105	12	6	1	1	2	0	0	18	6	0	3	14	3	6	0	0	0	24
-Observation and terpretation	5	Physician with	46	114	3	25	4	4	5	3	4	0	0	3	4	7	19	18	1	1	0	0
		Other	45	99	5	29	2	3	4	3	5	0	0	2	4	6	24	9	2	1	0	0
	4	Educator with	55	132	4	35	1	1	6	1	3	1	0	0	4	4	18	21	1	0	0	0
		Other	56	134	3	31	2	3	5	2	5	2	0	1	3	5	16	20	1	0	0	0
I-Problem-Solving	11	Physician with	61	153	4	15	2	4	15	5	4	1	0	4	14	6	3	20	2	0	0	1
		Other	56	143	4	16	2	1	17	3	1	2	2	2	16	9	3	16	3	1	0	2
	13	Educator with	56	129	5	16	2	0	23	0	2	1	0	0	22	5	2	18	3	0	0	1
		Other	56	123	6	13	2	2	18	2	2	1	1	2	16	8	3	19	3	0	0	1

* For definition of Categories see TABLE I.

Appendix 29

Chronological List of Joint Activities of American Board of Orthopaedic Surgery and Center for the Study of Medical Education in Relation to the Orthopaedic Training Study

Though the Center for the Study of Medical Education was the contractorial agent for the Orthopaedic Training Study, that Study was, from its inception, a joint effort of the Center and the American Board of Orthopaedic Surgery. Specifically, both cooperated fully in planning and staffing the research project, in addition to full and part time staff, the Board also recruited special consultants, one or more of whom was available for every stage of the study; the research staff held frequent, regularly scheduled meetings with both the full Board and its Examination Committee to review progress and to make joint decisions on detailed next steps; both the Board and the Center staff made regular reports to the entire orthopaedic community at the regional and national meetings of the association, as well as in their official journals and by means of specially prepared communications to each member of the specialty. In addition to these regular modes of cooperation and communications, the following special activities in which both Board members and Center staff participated should be noted:

November, 1962	Initial Request to the Center from the American Board of Orthopaedic Surgery for critical review of current examinations and for sugestions for improvement.
June, 1963	Preliminary analysis of the current certification progress and initial proposal for an extended joint study.
June, 1964	After a series of meetings betweenrepresentatives of the Center and the Board and beween them and the Public Health Service, the inaguration of the 4 year study with co-principal investigators, one from the Center and one from the Bard; the study purposes and plans explained by the Board to the orthopedic community and that community contacted by the Board, the Center and the Aerican Institutes of Research(to whom the Critical Icident Study had been sub-contracted) for participaion in that portion of the study.

December, 1964	Meeting with the Task Force on Process Analysis of Written Examinations; this Task Force was appointed by the Board to classify the written examinations by taxonomic levels
January, 1965	Observational analyses of oral examinations.
September, 1965	Meeting with the In-Training Examination Committee of the American Academy of Orthopaedic Surgeons to arrange cooperation with the Center and the Board in obtaining biographical information on residents and in developing and using new types of evaluation techniques in the regular In-Training Examinations for residents.
October, 1965	Special meeting with the Examination Committee of the Board to arrange for administering the new types of orals in the 1966 examinations for sponsorship of a training session for the oral examiners who would administer the experimental orals; and for administration of portions of the Written examination to examiners.
November, 1965	Meeting with a Task Force recruited by the Board to serve as a criterion group for scoring new types of written examinations.
December, 1965	Two-day training session for 40 examiners on administration and scoring of the new oral examinations.
April,1966	Establishment by the Board of an Office of Evaluation Services with a full-time director to serve as technical staff in discharging the Board's regular responsibilities for certification.
May, 1966	Special meeting with the Examination Committee of the Board for purposes of constucting a set of test specifications or blueprints which would define in detail the processes and content to be assessed in each certification examination.
June, 1966	Special meeting with the Board to arrange for appointment of a series of Task Forces composed of both Board and non-Board members to review

	and revise various recommendations for modification of certification procedures, to develop plans for implementing these recommendations, and to assist in developing the new materials they required.
August, 1966	Meeting with the In-Training Examination Committee of the Academy of Orthopaedic Surgeons to arrange for modifications in their examination procedures that would parallel the improvements being introduced in the Board examinations and to plan a three-day workshop on the construction of both written simulations and new types of multiple choice questions, conducted jointly by staff from the Center and from the newly established Office of Evaluation Service of the Board
September, 1966	Initial meeting with the Task Force appointed by the Board to develop systematic procedures for classification and updating of materials in the examination pool.
November, 1966	Initial meeting with the Task Force appointed by the Board to develop detailed plans for assessment of attitudes and skills.
November, 1966	Initial meeting with the Task Force appointed by the Board to establish a regular system for the development and review of new types of multiple choice questions that would insure a continual supply of good, new materials.
December, 1966	Initial meeting with the Task Force appointed by the Board to develop detailed recommendations for modifications in oral examination procedures such as to ensure that these examinations would be adequate to assess problem solving ability, interpretive skills, and certain professional attitudes.
December, 1966-- June, 1967	Series of meetings with sub-groups of the Task Force on Oral Examinations to develop standardized materials and directives for administering and scoring each of the new types of oral examinations recommended by the parent group
April, 1967	Initial meeting with the Task Force appointed by the Board to develop a system of scoring and weighting

	the various components of the certification examination so as to reflect the specifications established in the blueprint and to yield a profile of each candidate's performance that would indicate his achievement with respect to each critical requirement (not each examination technique.)
May, 1967	Meeting with the Chairmen of the several Task Forces to collate recommendations and materials from all working groups and to prepare for Board action a final set of recommendations regarding modifications of the certification system.
July, 1967	Meeting with the Board at which recommendations of the Task Force Chairmen were accepted and agreement was reached to waive the time and distribution requirements (for becoming Board eligible) for residents in selected experimental programs to be subsequently identified. (Note: This approval cleared the way for the extension of orthopaedic study in the new directions outlined in Section III, Chapter XII of the text.)
October, 1967	The first of a series of seven training programs designed to orient oral examiners in the technique of administering and scoring the new orals adopted for incorporation in the regular certification process (Note: These sessions held at various locations throughout the country were conducted jointly by the representatives of the Center staff and of the Board. One or more sessions was attended by each of the approximately 200 members of the cadre of oral examiners assembled to administer the 1968 examinations
November, 1967	Meeting with the Examination Committee of the Board at the first one to be called in accord with the newly established procedures for reviewing and selecting final review, codification, and selection of all materials for both oral and written components of the 1968 certification examinations, and for final determination of scoring procedures and minimal acceptable levels of performance on that examination.
January, 1968	Meeting with candidates and oral examiners following the administration of the first completely revised certification examination, toobtain their responses to a standardized questionnaire about the new system and their reactions to it in a series of group interviews.

May, 1968	Meeting with the Board at the first one called to review candidate performance after the initiation of the new system of profile scoring and reporting.
June--July, 1968	Conference with representatives of the Board on terminations of the first phase of the Orthopaedic Training Study and initiation of a second phase devoted to experimentation in curriculum and instructional methods.

Appendix 30

A Proposal for a Ten-year Follow-up

Hypotheses

The initial research design for the Orthopaedic Training Study provided for a 10-year follow-up to assess the predictive validity of the certification procedures of the American Board of Orthopaedic Surgery, in order to determine the extent to which these certifications procedures, as revised during the course of the initial study, are successful in differentiating candidates who subsequently become superior orthopaedic surgeons from those who do not. The long-term follow-up study outlined below is designed to investigate the following hypotheses:

1. That these are significatn differences between the certified and non-certified orthopaedic surgeons in the quality of patient care rendered by each;

2. That there are significant differences between the two groups with respect to the extent to which each participates in programs of continuation education;

3. That there are significant differences between the two groups with respect to the extent to which each assumes educational responsibilities for medical student and residency training; and

4. That there are significant differences between the two groups with respect to the extent to which each contributes to new knowledge in the field of orthopaedic surgery.

However, since certain of these hypotheses are either self-confirming (because speciality board certification is a prerequisite to participation in certain types of professional activities) or could be confirmed simply because a general factor 1 (i.e., intelligence and/or motivation) may explain part of the variance on any specific performance measure, the following additional hypothesis will be investigated in the proposed follow-up.

5. That the differences between the two groups (i.e. certified vs. non-certified) are relatively greater for candidates who applied for specialty board certification after the introduction of the revised certification procedures than for candidates who applied prior to the initiation of the Orthopaedic Training Study.

Methods

Sample Selection and Data Collection: In order to investigate these hypotheses it is proposed that the following samples be drawn from among the candidate pool who first applied for certification in January 1968, January 1969 and January 1970.* A sample of 100 candidates randomly selected from among those who were certified in the year in which they applied; a sample of 100 candidates randomly selected from those who failed at the time of their first application for Board certification but who were subsequently certified; and a third sample of 100 candidates randomly selected from among those candidates who by 1978 had not been certified. Analogous random samples are to be drawn from among candidates who first applied for Board certification in the years 1961, 1962 and 1963.**

It is proposed that in 1978 measures such as the following be obtained on each surviving member of the initial sample population of 600:

1. A self-report questionnaire indicating current sub-specialty, if any, current practice setting, current self-educational activities, current academic and research activities, interim publications, current professional associations, honors and the like;

2. A self-report patient log for a sample week to indicate the patient population served and the disposition of each patient seen.

3. (Except for chiefs of service within the sample population) a confidential rating from the chief of service indicating his evaluation of the unique pattern of competence of members of the sample population directly responsible to him;***

---

* January 1968 was the first year in which recommended revisions in certification procedures were fully implemented.

** These were the years immediately preceding the initiation of the Orthopaedic Training Study and would therefore be suitable for testing hypotheses (5) on preceding page.

*** The rating form would be designed to yield evidence on the aspects of competence derived from the Critical Incident Study that has guided the development of the new certifying procedures in this specialty; it is anticipated that this rating form would be closely similar to the recently developed Candidate Rating Form (see Appendix__) now in use by the Board.

In addition, for 25% of the cases randomly selected from each su-sample the following additional measures would be obtained:

1. A systematic review of hospital charts of patients each had admitted during a selected time period (probably 2 one-week samples) designed to obtain evidence on the quality of the diagnostic work-up and the management decisions of each, as revealed by the hospital records;

2. A half-day observation of each of the physicians in the sub-sample in his office contacts with patients; these observations to be systematically recorded an objective on a form designed for computer analysis.

In addition to the sample populations described above the following additional samples are to be drawn from candidates who first applied for certification in 1961 to 1963 and in 1968 to 1971. Five geographic centers will be selected in which there is a large group of practicing orthopaedic surgeons (for example, Boston, New York, Chicago); the sample population will consist of all those candidates currently practicing in the selected centers who first applied for specialty board certification in the specified years. This group of candidates will be subdivided into the same sub-samples as described above: i.e those who were certified at the time of their first application, those who were not certified at that time but who subsequently achieved specialty board certification prior to 1978 and those who by 1978 had not yet achieved specialty board certification. Within each geographic center all members of the sample population practicing in that center will be asked to complete a specially developed sociometric questionnaire designed to yield peer evaluations. The questions will be of a type* and of sufficient number both to disguise the evaluative nature of the questionnaire and at the same time to yield evidence suitable for ranking members of each sample with respect to major parameters of competence.

Analysis of Data

All data from questionnaires, self-reports, observations, peer and supervisor evaluations will be summarized in quantitative form and analyzed by analysis of variance techniques to determine the relative magnitude of within, and between, group variance. In addition, correlational and multipl regression analysis will be used to determine the relation between performance on the various components of the certifying procedure and indices of subsequent performance. These analyses, together with anecdotal material from the record review and observations, will be summarized to indicate the performance parameters, if any, on which the various sub-samples differ significantly from each other and to indicate the parameters, if any, on which the relative difference between sub-samples drawn from the 1968-1971 candidate pool is greater than that between sub-samples drawn from the 1961-1963 candidate pool.

---

* "To whom would you send a member of your family who developed speci kinds of orthopaedic problems?" "From whom do you normally first hear about a new development in your field?" "If you had a completely free choice whom would you like to have cover your practice if you had to be away for an extended period?" "If you had a son who wished to study orthopaedic surgery in this city with whom would you most like to have him train?"

Summary Comment

While the proposed study outlined above is designed primarily for purposes of gathering data on the predicted validity of alternative certification procedures, it will yield data regarding variations inpractice associated with age, practice setting, patient population served and the like; further these data can be used to identify the range and nature of problem encountered in the several types of practice settings, together with indication of the extent to which prior education and experience have prepared the clinician to deal with these problems optimally. These data should be of general significance in assisting program planners to improve professional education both the resident and post-graduate levels.